



SAPIENZA
UNIVERSITÀ DI ROMA

Il dialogo uomo-macchina: analisi di un *corpus* originale raccolto nell'ambito del progetto Speaky Acutattile

Facoltà di Lettere e Filosofia

Dottorato in Linguistica Storica e Storia Linguistica Italiana
XXVI Ciclo

Dipartimento di Scienze documentarie, linguistico-filologiche e geografiche

Addottorando

Dott. Fabio Poroli

Supervisor

Ing. Andrea Paoloni
Prof. Riccardo Gualdo

Anno Accademico 2013/2014

Il dialogo uomo-macchina: analisi di un corpus originale raccolto nell'ambito del progetto Speaky Acutattile

Indice

PREFAZIONE.....	6
1. INTRODUZIONE.....	8
1.1 I sistemi di dialogo, tra successo e sopravvalutazione.....	8
1.2 La ricerca linguistica e il dialogo uomo-macchina.....	12
2. I SISTEMI DI DIALOGO.....	14
2.1 Introduzione.....	14
2.2 Le interfacce.....	15
2.3 Architettura base di un sistema di dialogo.....	17
2.3.1 Automatic Speech Recognizer.....	19
2.3.2 Modulo di comprensione del linguaggio.....	21
2.3.3 Generazione del linguaggio naturale e sintesi vocale.....	24
2.4 Dialogue Manager.....	25
2.4.1 Gestione dell'iniziativa.....	26
2.4.1.1 L'iniziativa da un punto di vista teorico.....	26
2.4.1.2 L'iniziativa da un punto di vista applicativo.....	28
2.4.2 Architettura a stati finiti.....	29
2.4.3 Architettura frame-based.....	30
2.4.4 Gestione delle incomprensioni.....	32
2.4.5 La comprensione pragmatica.....	34
2.4.5.1 Dialogue acts.....	35
2.5 Siri.....	39
3. LE INTERAZIONI VERBALI TRA PARLANTI UMANI E SISTEMI DI DIALOGO AUTOMATICI.....	42
3.1 Introduzione.....	42
3.2 Situazioni comunicative e tipi di interazione tra umani e interfacce vocali.....	44
3.2.1 La situazione extralinguistica.....	44
3.2.2 Tipo d'interazione	46

3.3 La semplificazione nel dialogo uomo-macchina.....	48
3.3.1 Studi empirici.....	48
3.3.2 I modelli proposti per la semplificazione.....	51
3.3.2.1 Un sottocodice o una varietà con differenze funzionali.....	51
3.3.2.2 La rappresentazione dell'utente e la rappresentazione dell'interfaccia.....	52
3.4. Il linguaggio naturale.....	54
3.4.1 Human-like vs Tool-like.....	54
3.4.1.1 Tool-like.....	55
3.4.1.2 Human-like.....	56
3.4.2 La rappresentazione dell'utente.....	57
3.4.3 Interfaccia o interlocutore?.....	58
3.5 Sommario e conclusioni.....	61
 4. ANALISI DEL CORPUS C-ORAL-ROM.....	 62
4.1 Introduzione.....	62
4.2 Situazione e tipo di interazione.....	62
4.2.1 Situazione extra-linguistica.....	62
4.2.2 Struttura dell'interazione.....	63
4.3 Successo	65
4.4 Iniziativa.....	66
4.4.1 Gestione dell'iniziativa in C-ORAL-ROM.....	67
4.4.2 Posizione.....	67
4.4.3 Funzione.....	68
4.4.4 Tipologia delle funzioni.....	68
4.4.4.1 Backward-looking.....	68
4.4.4.2 Forward-looking.....	69
4.4.4.3 Chiusura.....	70
4.5 La gestione dell'iniziativa come risorsa per definire un profilo di interazione.....	70
4.6 Analisi.....	71
4.6.1 Dialoghi senza iniziativa.....	71
4.6.2 Dialoghi con iniziativa.....	75
4.7 Conclusioni.....	79
 5. IL SISTEMA SIMULATO PER LA RACCOLTA DEL CORPUS SPEAKY	 81
5.1 Introduzione.....	81
5.2 La tecnica del Mago di Oz.....	82
5.3 Variabili.....	83
5.4 Fasi della raccolta.....	84

5.5 Critiche e limiti.....	85
5.6 La raccolta del corpus Speaky.....	86
5.6.1 Il progetto Speaky.....	86
5.6.2 L'architettura dell'interfaccia vocale.....	87
5.6.3 La raccolta del corpus Speaky.....	89
5.6.3.1 Ambiente e situazione extralinguistica.....	89
5.6.3.2 Soggetti.....	90
5.6.3.3 Scenari.....	90
5.6.3.4 Compiti.....	92
5.6.3.5 Simulazione del sistema.....	94
 6. ANALISI DEL CORPUS SPEAKY.....	 99
6.1 Introduzione.....	99
6.2 Caratteristiche del corpus.....	100
6.2.1 Criteri della trascrizione.....	103
6.2.2 Un esempio di dialogo con il sistema.....	105
6.3 La gestione del turno nel corpus Speaky.....	107
6.3.1 Introduzione.....	107
6.3.2 Architettura dei turni e gestione dell'iniziativa.....	108
6.3.2.1 Architettura dei turni del sistema nel corpus Speaky.....	109
6.3.2.2 Saluti di apertura.....	113
6.3.2.3 Saluti di chiusura.....	116
6.3.2.4 Feedback di completamento.....	120
6.3.3 La gestione del turno.....	127
6.3.3.1 La gestione del turno nella conversazione uomo-uomo.....	127
6.3.3.2 La gestione del turno nelle interfacce grafiche.....	129
6.3.3.3 La gestione del turno nelle interfacce vocali.....	130
6.3.3.4 Voice Activity Detection.....	131
6.3.3.5 Rilevamento della cessione del turno.....	132
6.3.3.6 La gestione del turno nel corpus Speaky.....	134
6.3.4 Sommario e conclusioni.....	143
6.4 La gestione degli errori nel corpus Speaky.....	145
6.4.1 Introduzione.....	145
6.4.2 Tipologie d'errore nell'interazione con un sistema automatico.....	147
6.4.3 Strategie di gestione degli errori.....	150
6.4.4 Simulazione degli errori	155
6.4.4.1 Simulazione degli errori nel corpus Speaky.	156
6.4.5. Situazioni di mancata comprensione.....	158
6.4.5.1 Sequenza di riparazione.	159
6.4.5.2 Ripetizioni e riformulazioni.....	160
6.4.5.3 Ripetizioni e riformulazioni nel corpus Speaky.....	161
6.4.6. Situazioni di errata acquisizione.....	186
6.4.6.1 Sequenza di riparazione.....	187
6.4.6.2 Individuazione dell'errore.....	191
6.4.6.3 Strategie di correzione a confronto.....	192
6.4.6.4 Proprietà della correzione.....	195
6.4.6.4.1 Riprogrammazioni del compito.....	195

6.4.7 Conclusioni.....	202
7. CONCLUSIONI.....	204
8. BIBLIOGRAFIA.....	208

Prefazione

Il lavoro è incentrato sull'analisi di alcuni aspetti delle interazioni tra persone e sistemi di dialogo automatici (interfacce artificiali in grado di interagire verbalmente con una persona allo scopo di svolgere una determinata attività, cfr. capitolo 2) ed è stato condotto nell'ambito del progetto Speaky Acutattile, una piattaforma di assistenza per anziani e non vedenti, dotata di interfaccia vocale.

Nonostante l'uso delle interfacce vocali sia ormai in espansione anche in Italia (inizialmente nell'automatizzazione dei centralini telefonici, successivamente in parallelo al successo degli smarritone, spesso dotati di interfaccia vocale, tra tutti Siro dell'appellazione), la ricerca nell'ambito del dialogo uomo-macchina è stata molto poco battuta per quanto riguarda la lingua italiana. Al dialogo uomo-macchina è dedicata una parte del volume collettaneo sul dialogo Bazzanella 2002: la sezione è composta da due contributi introduttivi al tema (uno incentrato sui sistemi, Damiano 2002, un altro più prettamente linguistico, Danieli 2002); mentre il resto della letteratura scientifica è limitato a pochi altri contributi di carattere linguistico (Bazzanella et al. 1999, Danieli 2004).

L'ambito è stato poco esplorato dalla linguistica italiana soprattutto per via della mancanza di progetti industriali, fondamentali per catalizzare interesse e finanziamenti. Per il medesimo motivo, il panorama è caratterizzato da una sostanziale scarsità di dati a disposizione, limitati a due soli *corpora* di dialogo uomo-macchina, raccolti nell'ambito del progetto CLIPS e del progetto C-ORAL-ROM. L'inclusione di *corpora* di dialogo uomo-macchina in questi progetti non ha portato, tuttavia, a una maggiore attenzione sul tema né ha dato seguito ad alcun contributo scientifico che sfruttasse i dati a disposizione.

Per questo, la prima parte della tesi sarà dedicata a descrivere sommariamente i sistemi di dialogo e il dialogo uomo-macchina. Nell'introduzione si approcceranno le interfacce vocali come fenomeno sociale, contestualizzando alcuni “miti” mediatici e collocando la linguistica all'interno del processo di progettazione e

sviluppo di un'interfaccia vocale. Il capitolo 2 e il capitolo 3 saranno dedicati rispettivamente ai sistemi di dialogo (stato dell'arte e funzionamento) e ai lavori sul dialogo uomo-macchina, condotti principalmente all'estero. Nel capitolo 4 verrà presentata l'indagine condotta sul *corpus* C-ORAL-ROM (capitolo 4), un *corpus* esiguo ma particolarmente pertinente per illustrare l'interazione con un sistema di dialogo semplice programmato per svolgere un solo tipo d'attività (informazioni sugli orari ferroviari), che tuttavia consente all'utente di controllare l'interazione (iniziativa mista). Lo studio è stato condotto nell'ambito della mia visita al Laboratorio di Linguistica Informatica dell'Universidad Autonoma de Madrid. I capitoli seguenti (capitoli 5 e 6) costituiscono invece il tema centrale del progetto di ricerca, ovvero la raccolta e l'analisi del *corpus* per la progettazione di Speaker: nel capitolo 5 verrà illustrato il metodo di raccolta: il *corpus* infatti non è stato ottenuto con un sistema reale ma con un sistema simulato, tramite la tecnica del Mago di Oz, di cui si riporteranno le specifiche e un modello di gestione dell'iniziativa. Nel capitolo successivo verranno analizzati due momenti critici del dialogo uomo-macchina: la situazione d'errore e il passaggio del turno.

1. Introduzione

1.1 I sistemi di dialogo, tra successo e sopravvalutazione

Le interfacce che consentono l'interazione con un hardware o un software tramite linguaggio naturale, benché abbiano una storia quasi cinquantennale¹, sono diventate solo recentemente un argomento di attualità, grazie all'attenzione dedicatagli in questi ultimi anni dalle due grandi multinazionali della tecnologia: Apple e Google. La prima ha acquistato qualche anno fa la società Siri, proprietaria dell'omonimo sistema di dialogo per smartphone, e ha successivamente implementato ogni iPhone, (a partire dalla versione 4S) e iPad di tale interfaccia vocale; la seconda ha da poco lanciato i Google Glass, uno strumento, ancora in fase di test, indossabile come un paio di occhiali e adoperabile principalmente tramite un'interfaccia vocale. Gli investimenti delle due grandi aziende hanno portato molta attenzione mediatica: nelle pubblicità e negli articoli su riviste per appassionati di tecnologia l'interazione vocale con sistemi artificiali (intesi principalmente come smartphone, tablet e pc) è descritta come un necessario e naturale approdo per l'interazione uomo-macchina, dopo la “lunga” fase delle interfacce grafiche, iniziata con il primo Windows e arrivata agli attuali *touch screen*.

Mettendo per il momento da parte le limitazioni tecnologiche, la superiorità delle interfacce vocali rispetto alle equivalenti interfacce grafiche è fondata su alcune caratteristiche. Innanzitutto, sono più semplici da usare: l'uso del linguaggio naturale come strumento d'interazione non obbliga gli utenti a dover apprendere né un linguaggio né un metodo di interazione diverso da quello quotidiano: sono quindi potenzialmente usabili da tutti. Inoltre, se dotate di un'architettura a iniziativa mista², permettono di svolgere un'attività senza dover seguire una sequenza predefinita e costrittiva di passaggi (ad esempio, per la prenotazione online di un biglietto del

¹ La prima interfaccia per l'interazione con un sistema artificiale tramite l'uso del linguaggio naturale (per via scritta), ELIZA, fu presentata nel 1966 (Weizenbaum, 1966).

² Termine che verrà ripreso e approfondito in seguito (2.4.1): basti intanto intendere l'iniziativa mista come la possibilità, per l'utente, di non dover dipendere unicamente dal precedente turno del sistema artificiale, e di poter condizionare la successione dei turni a seconda della strategia usata per l'attività.

treno), i quali, infatti, possono essere negoziati durante l'interazione proprio come se si stesse conversando con una persona. Consentono, inoltre, di non dover usare le mani o guardare uno schermo: un vantaggio importante in situazioni che non permettono l'uso delle equivalenti interfacce grafiche (ad esempio, in automobile) o per aiutare persone, come non vedenti o disabili agli arti superiori, che di fatto non possono interagire con strumenti digitali per via visiva e tattile. Insomma, le interfacce in linguaggio naturale sono generalmente viste come più veloci, meno costrittive e adattabili a un grande numero di situazioni e persone.

I vantaggi appena descritti sono piuttosto ottimistici e non rispecchiano del tutto la realtà attuale, condizionata da ulteriori limitazioni tecnologiche. Se è vero che i riconoscitori vocali sono molto migliorati negli ultimi anni (requisito essenziale per il successo delle interfacce), soprattutto grazie alla possibilità di sfruttare grandi quantità di dati per perfezionare i modelli statistici, e se è vero che le reti mobili ne hanno reso possibile l'uso senza dover possedere materialmente grandi database sui propri smartphone o tablet, le difficoltà maggiori risiedono su altri piani del processo di interazione con un sistema, soprattutto a livello pragmatico e semantico. Attualmente, nessun sistema consente pienamente l'iniziativa mista, se non per singole e semplici attività, in cui eventuali errori sono poco importanti (e risolvibili velocemente) e che prevedono poche variabili da impostare (come ad esempio una sveglia). Anche la gestione del turno è, ad oggi, tutt'altro che naturale, essendo il più delle volte delegata ad azioni fisiche, come la pressione di un pulsante (*push-to-talk*), e può risultare molto complessa per quei soggetti che hanno poca dimestichezza con gli strumenti digitali e che costituiscono uno dei principali target delle interfacce vocali.

Pur possedendo tali limiti, il sistema di dialogo Siri (cfr. 2.5), essendo un'ottima interfaccia per alcune applicazioni molto semplici (impostare una sveglia, un promemoria o mandare un sms) ha spostato in avanti i requisiti necessari per un sistema dialogo commerciale. Tuttavia, contrariamente alla parallela interfaccia grafica che controlla ogni applicazione dell'appellazione, Siri non è riuscito ancora a riscontrare successo nell'uso e a risultare, per l'utenza, un'interfaccia che migliori l'interazione. Purtroppo non ci sono molte indagini sul reale uso delle interfacce

vocali rispetto alle equivalenti interfacce grafiche, condotte su campioni rappresentativi: una di queste rileva che degli utenti americani dell'ultimo sistema operativo Apple iOS7 solo il 15% ha usato Siri, e tra questi almeno la metà lo giudica poco utile ed eccessivamente sopravvalutato (McPheat 2013). Va inoltre rilevata la frequenza di articoli o post, sui siti di appassionati di tecnologia, che ruotano attorno alla domanda: “Chi di voi usa realmente Siri?”.³ Gli utenti usano Siri principalmente per le funzioni più semplici (messaggi, sveglia, telefonate e promemoria), ma rimangono comunque pochi rispetto a quelli che, anche per le suddette applicazioni, usano l'alternativa interfaccia touch screen, che segue l'attuale standard per l'interazione con gli smartphone e i tablet, e che consente di usare tutte le applicazioni, tra cui alcune decisamente importanti, come la navigazione su internet. Proprio la navigazione sul web costituisce un secondo grande problema con cui dovranno confrontarsi le interfacce vocali: la fruizione della mole di informazioni quotidiana (tramite web, social network e applicazioni di ogni tipo) avviene oggi principalmente tramite un canale visivo: questo ha condizionato negli anni la stessa modalità di lettura di un testo, che procede a salti e individuazione di parti importanti senza che necessariamente venga letta singolarmente ogni parte. Anche se le interfacce vocali fossero perfette nella comprensione, la ricezione visiva, in questa situazione comunicativa, rimarrà ancora per molto più pratica e veloce. Infine, bisogna aggiungere che gli “analfabeti digitali”, persone che incontrano molte difficoltà nell'uso di interfacce grafiche e che costituiscono uno dei consueti target nei progetti di interfacce vocali, sono in calo, visto il successo di massa di computer e smartphone (oltre che per evidenti ragioni legate all'età e al ricambio generazionale).

La domanda posta ai possessori di iPhone (“Chi di voi usa realmente Siri?”) è frutto di considerazioni molto pratiche, ma che chiunque lavori in questo ambito dovrebbe porsi, e che, come vedremo, toccano anche le altre discipline che possono contribuire alla progettazione di interfacce vocali, tra cui la linguistica. In quali situazioni serve un'interfaccia vocale? Per quali applicazioni? Al di là dell'evidente

³ Naturalmente gli appassionati sono un target poco indicativo per il successo di una nuova tecnologia, costituendo di fatto una piccola fetta dell'utenza decisamente sbilanciata verso la curiosità e la sperimentazione di nuove interfacce.

gap di complessità tra un'interfaccia grafica e una vocale, la seconda migliorerà realmente l'interazione? La velocizzerà? Insomma, le interfacce vocali possono essere migliori di quelle grafiche? In questo caso con “migliori” intendo “più efficienti per svolgere un'attività”, come ad esempio la semplicità d'uso delle interfacce grafiche (Windows) rispetto alle precedenti interfacce con stringhe di comando (DOS), che richiedevano invece una grande competenza tecnica per l'interazione. Da questa prospettiva strumentale, purtroppo non ancora supportata dai dati, Siri appare come un'interfaccia senz'altro evoluta e capace di agire tramite comandi in linguaggio naturale, ma che tuttavia rimane secondaria rispetto all'equivalente interfaccia grafica, meno futuristica e più efficiente per ognuna delle applicazioni del telefono. Non è un caso se gran parte degli elogi degli utenti riguardino l’“umanità” del sistema, prima ancora della sua efficienza. Al di là di un riconoscitore vocale molto buono su un hardware all'avanguardia nel mercato, Siri deve infatti molta della sua fama a quelli che possiamo definire gli “specchietti per allodole” dei sistemi di dialogo, come risposte a domande sul senso della vita, a insulti o a inviti a cena⁴.

1.2 La ricerca linguistica e il dialogo uomo-macchina

Quanto detto finora potrebbe non essere l'inizio ideale per un lavoro sull'interazione vocale uomo-macchina: abbiamo evidenziato come il successo, ad oggi, sia in gran parte mediatico; come l'interazione con i sistemi artificiali sia volta, oggi, verso un'altra direzione (ovvero una ricezione-produzione principalmente scritta e visiva); come le interfacce vocali, sia per limitazioni tecnologiche sia per la presenza di una valida alternativa grafica, siano ancora molto lontane dall'essere interfacce universali (come ad esempio è in poco tempo diventato il touch screen, affine alle interfacce grafiche precedenti, permettendo un'interazione ancora meno costrittiva).

Tale critica alla tecnologia vocale è tuttavia funzionale a sgombrare il campo

⁴ Vale la pena notare come Siri sia molto più conosciuto della equivalente interfaccia vocale di Google, su Android, condividendone, tuttavia, quasi tutte le caratteristiche e usando un riconoscitore ugualmente buono. Al di là del maggiore prestigio dell'iPhone, bisogna segnalare che l'interfaccia vocale di Google non è “umanizzata”.

da alcuni pregiudizi, che finiscono per condizionare anche l'apporto che la linguistica (e altre discipline sociali e umane, come l'antropologia) può dare nella progettazione di sistemi di dialogo: l'idea, ad esempio, che l'obiettivo ultimo della ricerca nel campo dei sistemi di dialogo sia un interlocutore artificiale con competenze uguali o simili a quelle umane. Se l'obiettivo di un'interfaccia vocale non è chiaro per chi, come un linguista, è esterno in gran parte del processo di produzione e lavora direttamente sul rapporto tra il sistema e l'utente, la ricerca rischia di essere poco pertinente sia da un punto di vista applicativo sia da un punto di vista teorico. La celebre massima di Jelinek «Every time I fire a linguist, the performance of the speech recognizer goes up»⁵, al di là delle possibili strumentalizzazioni, andrebbe intesa come uno stimolo in senso strumentale: la linguistica può dare apporto in ambiti applicativi che le competono, come i sistemi di dialogo, a patto di puntare verso l'efficienza e la semplicità dell'interfaccia, prima ancora che la sua “completezza” linguistica, da avvicinare a quella di un essere umano. Come vedremo nel capitolo 3, i problemi possono nascere nel momento in cui un sistema di dialogo è visto come un interlocutore da perfezionare e migliorare, piuttosto che un'interfaccia che sfrutta la metafora dell'interlocutore per semplificare l'interazione, così come le interfacce grafiche sfruttano la metafora spaziale (il desktop).

I sistemi di dialogo non devono essere messi a confronto con gli interlocutori umani, ma con altre interfacce; ad oggi, devono quindi saper svolgere le medesime attività in maniera più semplice e veloce delle equivalenti interfacce grafiche. Ciò non riduce l'apporto che possono dare le diverse branche della linguistica alla progettazione e allo sviluppo, ma lo seleziona e lo regolarizza. La linguistica deve lavorare nell'ambito dell'usabilità e degli *user studies*: non partendo dal dialogo tra persone ma studiando ciò che può migliorare l'esperienza con un'interfaccia da un punto di vista linguistico. Il dialogo tra persone può essere usato come termine di paragone solo quando una caratteristica dell'interfaccia è particolarmente distante dall'uso

⁵ Per quanto molto conosciuta, la frase manca di un preciso riferimento ed è conosciuta anche in un'altra forma «Anytime a linguist leaves the group the recognition rate goes up» (Jurafsky – Martin 2009). Young 2010 ha provato a contestualizzarla: un linguista si licenziò dal team e Jelinek lo rimpiazzò con un'ingegnere, notando successivamente che i risultati del riconoscitore cominciavano a migliorare; così provò a spostare un altro linguista del suo team in un altro settore, con lo scopo di rimpiazzarlo con un altro ingegnere.

quotidiano del mezzo linguistico, e costituisce un serio limite all'uso di tale mezzo per semplificare l'interazione⁶.

Come vedremo nel capitolo 3, una volta riconosciuta da numerosi studi la semplificazione di funzioni e strutture linguistiche nell'interazione con un agente non umano, ha poco senso concentrarsi su migliorie che raramente possono essere sfruttate dagli attuali sistemi, né tantomeno dalle persone che interagiscono con tali sistemi (come ad esempio l'uso dei segnali discorsivi, molto rarificato nell'interazione con un sistema). È necessario, piuttosto, partire da studi empirici su *corpora* di dialogo uomo-macchina, già “asciugati” dalle altre funzioni normalmente non coinvolte nell'interazione con un sistema.

⁶ Ad esempio, la gestione del turno, attualmente molto costrittiva e molto distante da quella usualmente negoziata tra persone. Cfr. 6.3.3.3

2. I sistemi di dialogo

2.1 Introduzione

Nella situazione comunicativa di dialogo uomo-macchina uno dei partecipanti è un sistema automatico, programmato per interagire verbalmente con una persona allo scopo di svolgere un determinato compito. A differenza dello studio delle situazioni comunicative quotidiane, per lo studio del dialogo uomo-macchina si ha la possibilità, da un punto di vista linguistico (ma anche antropologico, sociologico, psicologico, ecc.), di descrivere esaustivamente le caratteristiche del partecipante automatizzato: come funziona, cosa considera rilevante, come si comporta in determinate situazioni; tali informazioni, come ben sa chi si occupa di dialogo, sono più complesse da valutare per quel che riguarda il parlante umano. Inoltre, anche il tipo di dialogo che caratterizza tali interazioni è nettamente delimitato dagli obiettivi dell'utente e dalle possibilità del sistema (orientate a svolgere un determinato compito). Tali dialoghi sono chiamati generalmente “dialoghi pratici” (Allen et al. 2000, Leech – Weisser 2005), per via del forte orientamento verso la risoluzione di un compito (come la richiesta di informazioni o l'acquisto di un prodotto) e dalla delimitazione netta dell'argomento (o dominio) su cui vertono (cfr. 3.2). Pur avendo questo minimo comune denominatore, i sistemi di dialogo si contraddistinguono per una grande varietà di caratteristiche e architetture diverse, che incidono sulla *facies* dell'interazione. Come fa notare Danieli 2002:

«Se i sistemi automatici basati sull'interazione vocale sono ormai largamente diffusi per servizi accessibili usando il telefono, fa però torto alla realtà dei fatti considerare alla stregua di sistemi di dialogo le applicazioni che consentono al parlante di pronunciare una o due parole per turno. Nell'uso della parola *dialogo* nelle applicazioni correnti possiamo tracciare un *continuum* ai cui estremi collochiamo da una parte il dialogo persona-macchina, inteso come scambio effettivo e

negoziiazione di contenuti tra un parlante e un agente artificiale, dall'altra i cosiddetti sistemi IVR (*Interactive Voice Response*) che chiedono al parlante di scegliere una parola da un insieme proposto. All'interno di questo *continuum* spiccano applicazioni che possiedono, in misura maggiore o minore, tratti caratteristici del dialogo persona-macchina quali la capacità di mantenere aggiornato il contesto linguistico e referenziale dell'interazione, l'abilità di interpretare frasi del parlante che non siano esplicitamente sollecitate dall'agente artificiale e la capacità di risolvere fraintendimenti dovuti a errori del riconoscimento acustico.»

Ciò che accomuna sistemi così diversi, più che la pertinenza o meno della parola “dialogo”, è la loro natura di interfacce. Il rischio che si corre nel guardare ai sistemi di dialogo come interlocutori ancora poco sviluppati (e solo per tale motivo interfacce) è dimenticare il ruolo primario del sistema: la capacità di svolgere un compito in maniera più semplice e immediata di altre interfacce. Come osserva Dahlbäck 1991:

«The goals of research on dialogue in computational linguistics such as “Getting computers to talk like you and me” or developing interfaces “that will allow the user to forget that he is questioning a machine”, are not only difficult to reach. They are misconceived.»

2.2 Le interfacce

Le interfacce consentono l'interazione tra uomini e strumenti più o meno complessi: in altre parole sono sistemi il cui ruolo è permettere agli utenti di usare oggetti e sfruttarne le funzionalità, senza per questo doverne imparare il funzionamento. L'interfaccia può essere tutt'uno con lo strumento (ad esempio, un paio di forbici), il suo uso può essere immediato e “analogico” con l'uso dello strumento (ad esempio, il volante di una macchina); altrimenti, nel caso di strumenti più complessi, è costruita su una “metafora”: il ruolo della metafora è quello di

inquadrare un'entità il cui funzionamento è sconosciuto all'interno di uno spazio che l'utente già conosce o su cui può operare facilmente (Norman 1988). L'esempio di interfaccia metaforica più comune, con cui abbiamo a che fare ogni giorno è l'ambiente desktop: usiamo con naturalezza il personal computer, tramite il mouse (e la tastiera), come se fosse uno spazio fisico entro cui possiamo spostare oggetti, eliminarli, ecc. Le interfacce grafiche sono ormai molto usate e fanno parte della nostra vita quotidiana; soprattutto per le nuove generazioni è difficile pensare a qualcuno che non sappia usare un personal computer; tuttavia, l'artificiosità di tale interfaccia è evidente quando proviamo a insegnare l'uso del personal computer a una persona anziana: ciò che oggi ci sembra assolutamente naturale e immediato, richiede infatti un processo di apprendimento di numerose regole e convenzioni tutt'altro che naturali.

Le interfacce vocali ambiscono, invece, alla “naturalezza” vera e propria, basandosi su un insieme di regole e convenzioni che tutti conosciamo: quelle della comunicazione nella nostra lingua. Oltre a questo obiettivo, su cui c'è ancora molto da lavorare, le interfacce vocali hanno ulteriori vantaggi rispetto alle interfacce grafiche. Un input vocale⁷ pronunciato con lo scopo di eseguire una determinata azione è senz'altro più semplice e immediato dell'esplorazione attraverso cartelle e menu; l'interfaccia grafica obbliga, infatti, gli utenti a dover seguire quello che propone il sistema e a doversi conformare all'interfaccia. L'obiettivo delle interfacce vocali è, al contrario, quello di sollevare l'utente dall'impegno di capire il modo in cui un progettista ha programmato un sistema. In aggiunta, il canale vocale è fondamentale per le persone affette da disabilità (come la cecità o l'impossibilità di usare gli arti superiori) e molto comodo per le numerose situazioni in cui si hanno le mani occupate (come ad esempio in macchina). Allo stato dell'arte attuale le interfacce vocali sono ancora a un livello non sufficiente da sostituire le interfacce grafiche: la difficoltà di comprendere il linguaggio naturale rende le interazioni spesso frustranti, annullando di fatto i vantaggi di un tale approccio; tuttavia, soprattutto grazie alla diffusione di strumenti come lo smartphone le aziende hanno cominciato a investire sempre di più sulle interfacce vocali: un esempio concreto è il

⁷ Quando usati, e a meno di altre indicazioni, *input* e *output* sono da intendersi, rispettivamente, come parlato rivolto al sistema e come parlato proveniente dal sistema.

successo di Siri, il sistema di dialogo di serie su uno degli smartphone più venduti al mondo, largamente usato soprattutto da persone non vedenti (cfr. 2.5).

2.3 Architettura base di un sistema di dialogo

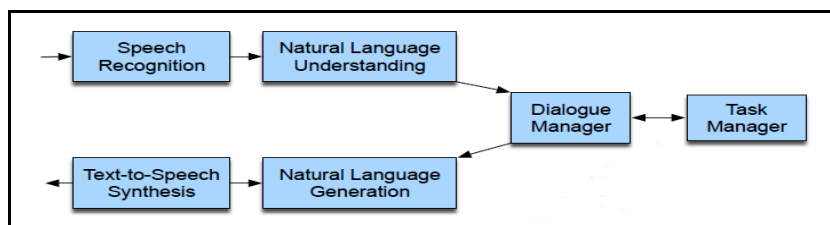


Fig. 1: Architettura base di un sistema di dialogo (da Jurafsky – Martin 2009)

La progettazione dei sistemi di dialogo coinvolge diversi rami della linguistica computazionale: il riconoscimento acustico, l'interpretazione semantica, la sintesi vocale, l'*information retrieval*. Un sistema di dialogo è composto, infatti, da più moduli, le cui operazioni sono strettamente legate all'architettura del sistema, definita *in primis* dal modulo centrale: il *dialogue manager* ('gestore di dialogo'), il quale, oltre a definire cosa far dire al sistema in relazione a quanto detto dal parlante, si interfaccia con il programma da usare. Generalmente, i moduli che compongono l'architettura base di un sistema sono il riconoscitore vocale, il modulo di comprensione del linguaggio, il dialogue manager (generalmente associato a un *task manager*, il modulo che rende effettive le azioni volute dal parlante), il generatore del linguaggio e la sintesi vocale (come si può vedere nella fig. 1).⁸

L'enunciato del parlante viene decodificato dal riconoscitore vocale in una stringa testuale successivamente interpretata dal modulo di comprensione; la rappresentazione semantica viene quindi inviata al dialogue manager, che si interfaccia con il task manager per eseguire l'azione e decide come comportarsi a livello verbale: se segnalare il successo dell'azione, se chiedere ulteriori informazioni per eseguire l'azione, se chiedere conferma o una ripetizione di quanto detto; la

⁸ Per una visione generale dei sistemi di dialogo cfr. Varile – Zampolli 1997: 199-222, De Mori 1998: 523-562; Cassell 2009, Pieraccini 2012.

risposta del sistema viene quindi generata dai due moduli di produzione del linguaggio. Il dialogue manager è senz'altro il modulo più importante e specifico dei sistemi di dialogo, in quanto responsabile delle scelte fatte dopo ogni turno del parlante, facendo sempre riferimento a un protocollo globale che ne definisce il comportamento in ogni situazione⁹. Tutti gli altri moduli sono declinati e tarati sulla base di come viene condotta l'interazione dal dialogue manager, come vedremo a proposito della gestione dell'iniziativa, degli errori, e della formulazione delle domande. Prima di andare a vedere la gestione del dialogo, verrà descritto brevemente ogni modulo, pur tenendo sempre presente il contesto dialogico e la dipendenza dal modulo più importante.

2.3.1 Automatic Speech Recognizer

Il modulo ASR (*Automatic Speech Recognizer*, 'riconoscitore vocale automatico') si occupa della digitalizzazione e della decodifica del segnale vocale da parte del parlante: il risultato viene successivamente inviato al modulo di comprensione del linguaggio. Il riconoscimento vocale è tradizionalmente associato, soprattutto a livello commerciale, alla trascrizione automatica del parlato: compito di per sé più complesso rispetto a quello che si trova a svolgere nell'architettura di un sistema di dialogo. Nei sistemi di dettatura il riconoscitore vocale deve infatti gestire un vocabolario molto più grande e non limitato a un solo dominio. D'altra parte, il riconoscitore vocale di un sistema di dialogo, pur gestendo un vocabolario più piccolo, deve lavorare in una situazione comunicativa meno controllata rispetto alla dettatura e confrontarsi quindi con i tradizionali fenomeni associati al parlato spontaneo.

I riconoscitori vocali si suddividono tra *speaker dependent* e *speaker independent*: i primi necessitano di una fase di addestramento, nella quale il riconoscitore, già addestrato su *corpora* di grandi dimensioni, si tara sulle caratteristiche vocali dell'utente (alcuni riconoscitori sono programmati per

⁹ Se è vero che il protocollo è fisso, i sistemi progettati in un'ottica *speaker dependent* (cfr. 2.3.1) si addestrano e adattano sull'utente: oltre a migliorare i risultati (apprendendo le abitudini dell'utente e comportandosi di conseguenza), miglioramento che di fatto coincide con un cambiamento nella gestione del dialogo, alcuni sistemi sono programmati per riuscire i termini che un certo utente normalmente usa (cfr. 2.3.3).

continuare a calibrarsi sull'uso del parlante anche dopo la prima fase); i secondi, essendo rivolti a più utenti e dovendo interpretare diverse caratteristiche vocali, vengono addestrati solo su *corpora* di grandi dimensioni. La scelta tra i due tipi di riconoscitore viene fatta in relazione al target del sistema: servizi come call center automatici sono *speaker independent*; al contrario, un sistema di dialogo su uno smartphone (come SIRI) o sul proprio personal computer può essere addestrato dallo stesso utente.

L'ASR viene addestrato con numerose registrazioni di ogni fonema: da questo addestramento vengono ricavati determinati valori (*feature number*) con una determinata probabilità di ricorrere a seconda del fonema. Durante la fase di riconoscimento, ogni stringa di *feature numbers* osservati assegna un valore di probabilità per ogni fonema, che viene ricalcolata insieme agli altri fonemi e sovrapposta al vocabolario del sistema, per ottenere la parola più probabile. A questo valore estratto, si aggiunge un ulteriore valore, sintattico, calcolato sempre su basi statistiche dedotte da *corpora* e che determina qual è la probabilità di una parola di succedere ad altre parole osservate. Questo modello è chiamato “a N-grammi”: ogni parola ha una determinata probabilità di comparire dopo un'altra parola, o dopo una sequenza di più parole (in genere tre o quattro). L'incrocio tra le caratteristiche acustiche di una traccia vocale e le possibilità di composizione (addestrate statisticamente) permettono al sistema di identificare stringhe di testo con alti tassi di successo.

Bisogna precisare che questi modelli probabilistici vengono applicati principalmente ai sistemi di dettatura, che possiedono un lessico indipendente dal dominio specifico, presentano un vocabolario molto grande (tra le 20000 e le 60000 parole) e sono addestrati su *corpora* abbastanza grandi da permettere la rilevanza di analisi statistiche. Inoltre, un ASR generico di dettatura si trova a operare in contesti di enunciazione diversi da quello in cui deve lavorare il modulo di riconoscimento di un sistema di dialogo: nella dettatura o nella lettura di un testo le percentuali di errore sono molto basse (non oltre il 3%), percentuali che si alzano per il parlato non letto e nelle situazioni dialogiche spontanee (Jurafsky – Martin 2009).¹⁰

¹⁰ Anche se, come vedremo, la situazione comunicativa uomo-macchina comporta un parlato più controllato delle conversazioni spontanee.

La robustezza dei risultati che fornisce il riconoscitore può variare molto se lo stesso riconoscitore è integrato all'interno di un sistema di dialogo. Innanzitutto, la maggior parte dei sistemi di dialogo, anche se programmata per lavorare su più domini, possiede un vocabolario molto minore rispetto a un sistema di dettatura: questo rende più difficile un training statistico apposito (dovendo, nel caso, raccogliere *corpora* specifici di grandi dimensioni: operazione lunga e costosa). Inoltre, la situazione dialogica comporta vantaggi e svantaggi per l'ASR. I vari fenomeni linguistici ed extralinguistici (come esitazioni, false partenze, riformulazioni, pause piene, ecc.) che si possono presentare nell'interazione tra due parlanti e più generalmente nel parlato spontaneo non pianificato rendono più difficile il riconoscimento; allo stesso tempo, ogni azione conversazionale comporta uno spettro di reazioni, generalmente attinenti a quelle precedenti, sulla base del principio di cooperazione (Grice 1975): il parlato dell'utente è infatti condizionato dallo stato del dialogo. Di conseguenza, nella progettazione si può prevedere quale parte del vocabolario deve essere attiva a seconda dello stato della conversazione. Gli ASR integrati nei sistemi di dialogo possiedono quindi "grammatiche restrittive": a seconda dell'architettura prevista dal sistema, l'ASR viene costretto a riconoscere solo frasi compatibili con una struttura scritta a mano in fase di progettazione (grammatica a stati finiti, cfr. 2.3.2) o a considerare un ulteriore valore probabilistico calcolato a seconda dello stato del dialogo (grammatica statistica). Se ad esempio il sistema chiede la città di partenza per un volo, un ASR con una grammatica a stati finiti limiterà il vocabolario al solo riconoscimento delle città; allo stesso modo, nelle grammatiche statistiche, la probabilità calcolata dal riconoscitore sulla base della traccia acustica è influenzata dalle domande che sono state fatte dal sistema. All'interno di un sistema di dialogo l'ASR ha inoltre il compito di assegnare a ogni informazione acquisita un valore di certezza (*confidence*), per consentire al modulo di gestione del dialogo di giudicare l'affidabilità dell'interpretazione, sulla base del quale può decidere se acquisire l'informazione, se chiedere conferma di ciò che ha compreso o se richiedere la ripetizione (cfr. 2.4.4 e 6.4.2 per le strategie di gestione degli errori).

2.3.2 Modulo di comprensione del linguaggio

Il modulo di comprensione del linguaggio (spesso abbreviato con NLU, per ‘Natural language understanding’, o SLU, per ‘Spoken language understanding’, specificamente per il parlato) si occupa di interpretare il testo decodificato dal riconoscitore, allo scopo di fornirne una rappresentazione semantica adeguata al modulo di gestione del dialogo. Così come per il riconoscimento vocale, la comprensione automatica del linguaggio naturale ha numerose applicazioni al di fuori dei sistemi di dialogo automatici, tra cui l'estrazione di informazioni da *corpora*, la *automatic summarization* ('riassunto automatico') e la traduzione automatica (anche se per questa vengono usati, attualmente, approcci principalmente statistici).

Come abbiamo visto, il riconoscitore vocale dipende strettamente dal modulo di comprensione usato, che ne condiziona le grammatiche (dette, per questo, restrittive): l'ASR trascrive solo gli enunciati che sono rilevanti per il sistema, tutto ciò che è al di fuori del vocabolario previsto per il modulo NLU viene detto *out-of-grammar*. Il legame stretto tra i due moduli può essere visto nella grammatica chiamata *frame-and-slot* (Bobrow et al. 1977), molto usata nei sistemi di dialogo. Un *frame* è una classe di oggetti relativi al compito e al dominio su cui opera il sistema, a cui corrispondono determinati *slot* da riempire con valori chiave estratti dagli enunciati dell'utente. Prendiamo ad esempio la seguente richiesta di informazioni:¹¹

“Che treni ci sono da Roma a Milano il 27 aprile di mattina?”

Il sistema decodifica il testo usando un analizzatore semantico specifico e semplificato per il dominio (Lewin et al. 1999). L'analizzatore semantico estrae dal testo dei valori da associare alle entità semantiche specificate tra le regole fissate in fase di progettazione. A ogni entità semantica corrisponde una classe di termini, definita *a priori*:

RICHIESTA: informazioni, prenotazione, ecc.

¹¹ Riprendo l'esempio dal sistema simulato usato per il corpus Speaky (cfr. capitolo 5), i cui compiti sono stati formalizzati secondo una grammatica *frame-and-slot*.

ORIGINE: città, regione, stato, ecc.

DESTINAZIONE: città, regione, stato, ecc.

CITTÀ: Roma, Milano, Torino, ecc.

DATA: giorni dell'anno, domani, dopodomani, ecc.

ORARIO: una, due, tre, parte del giorno

PARTE DEL GIORNO: mattina, pomeriggio, sera, notte, ecc.

Sulla base di queste regole di comprensione, il modulo riempie gli slot necessari per l'esecuzione dell'azione:

RICHIESTA: informazioni treno

TRENO: origine, destinazione, data, orario, destinazione

ORIGINE: città

CITTÀ: Roma

DESTINAZIONE: città

CITTÀ: Milano

DATA: 27 aprile

ORARIO: ora, parte del giorno

PARTE DEL GIORNO: mattina

Questo tipo di grammatica è largamente usata dai sistemi commerciali, che generalmente svolgono compiti ben definiti su domini ristretti e facilmente formalizzabili. Tuttavia le regole sono strettamente legate al sistema, non sono generalizzabili e devono essere scritte “a mano” in fase di progettazione: un procedimento che su più compiti e domini può essere molto lungo. Inoltre, basandosi esclusivamente sul lessico trascritto dal sistema, la suddivisione in classi semantiche non permette di gestire le ambiguità e i casi in cui un termine può essere pertinente per più classi. Per questo alla grammatica semantica vengono spesso aggiunti modelli statistici, che assegnano a ogni parole chiave estratta una probabilità di comparire in una determinata richiesta, calcolata su dati precedentemente raccolti con lo stesso sistema. Altri approcci alla comprensione automatica sono basati su

modelli più genuinamente statistici, addestrati su *corpora* opportunamente etichettati: ad esempio, il modello di Pieraccini et al. 1991 si basa sulla probabilità che hanno determinate sequenze di parole estratte dall'enunciato di essere associate agli equivalenti ruoli semantici (o concetti) previsti dal sistema: la sequenza con maggiore probabilità viene quindi scelta per riempire gli slot relativi al compito.

Anche nel caso del modulo semantico, sia che abbia regole scritte a mano, sia che sia integrato con modelli statistici, vengono usate grammatiche restrittive di comprensione: a ogni stato del dialogo corrispondono determinate sequenze che possono essere interpretate e acquisite dal sistema.

2.3.3 Generazione del linguaggio naturale e sintesi vocale

Una volta compreso e contestualizzato l'enunciato, in base al compito che il sistema deve svolgere, spetta al modulo di generazione del linguaggio scegliere “cosa dire” e “come dirlo”. Il modulo si occupa di organizzare la forma con cui esporre all'utente i contenuti e gli intenti del sistema, in consonanza con le decisioni del dialogue manager. Anche in altri impieghi al di fuori dei sistemi di dialogo, la generazione del linguaggio naturale è collegata alla comprensione di un testo: in applicazioni come l'*information retrieval* o l'*automatic summarization*, il risultato che l'utente si aspetta deve essere organizzato secondo alcune regole che ne stabiliscono la rilevanza a seconda del testo decodificato.

Nella maggior parte dei sistemi di dialogo, soprattutto tra quelli che devono gestire su una sola applicazione e che operano su un dominio limitato, i contenuti vengono trasmessi all'utente attraverso la cosiddetta “generazione *template-based*”. L'enunciato del sistema segue una struttura sintattica fissa con delle variabili legate alle informazioni acquisite. Basandoci sull'esempio precedente: “A che ora vuoi partire da [ORIGINE:CITTÀ]?”, dove [ORIGINE:CITTÀ] è la città di origine che il sistema ha acquisito dall'enunciato del parlante. Il metodo è robusto e fortemente personalizzabile, ma ha come limite la portabilità su altri sistemi, domini e applicazioni, essendo strettamente legato a questi (come per l'architettura *frame-and-slot*) e necessitando di una scrittura a mano delle parti fisse delle sequenze.

Recentemente, nell'ambito della ricerca sulla generazione del linguaggio naturale, sono stato proposti modelli diversi e meno legati al singolo dominio e applicazione, esportabili perciò da un sistema all'altro, ma allo stesso tempo meno robusti e per cui è necessaria in ogni caso la personalizzazione delle regole per ogni applicazione (cfr. Reiter 2010). Innanzitutto, vengono rigidamente tenuti separati i compiti del modulo: la pianificazione dei contenuti, la pianificazione dell'enunciato e la sua realizzazione. La pianificazione del contenuto è la fase in cui il sistema, una volta compreso l'enunciato e in relazione a questo, decide cosa va detto in risposta all'utente: informazioni da fornire, feedback dell'esecuzione di un'azione, richieste di ulteriori dati rilevanti per completare il lavoro, richieste di ripetizione o conferma. Una volta decisa la mossa da fare, la seconda fase consiste nel pianificare linguisticamente (lessicalmente e sintatticamente) l'enunciato del sistema: vengono così selezionati i termini coerenti (in alcuni sistemi vengono usati gli stessi termini del parlante) e strutturati in un unico enunciato organizzato funzionalmente: per esempio, se il sistema chiede conferma di due informazioni che ha acquisito, invece di chiedere “Vuoi partire da Roma?” e “Vuoi partire di mattina?”, fonde le due domande in un solo enunciato “Vuoi partire da Roma di mattina?”. Infine il testo viene automaticamente formattato a seconda del modulo TTS (Text-To-Speech) e inviato al sintetizzatore vocale. A queste regole generali vengono talvolta aggiunti alcuni accorgimenti per rendere il dialogo meno ripetitivo e schematico: ad esempio, segnali discorsivi di feedback e presa del turno, o risposte con ellissi quando l'attività è ripetitiva, allo scopo di abbreviare i turni del sistema (i cosiddetti *tapered prompts*, Coehn et al. 2004). A prescindere da queste migliorie, la regola di massima della generazione del linguaggio naturale (applicato ai sistemi di dialogo) è che il sistema deve dire ciò che è in grado di comprendere: infatti, così come nel dialogo naturale gli interlocutori tendono ad adattare le proprie scelte lessicali a vicenda, allo stesso modo nel dialogo uomo-macchina il parlante tende ad adattare il proprio lessico a quello del sistema.

2.4 Dialogue Manager

Il gestore del dialogo controlla l'architettura e la struttura del sistema: acquisisce i dati dai moduli ASR e NLU, si interfaccia con le informazioni a disposizione e con le azioni che può eseguire, stabilisce quali contenuti trasmettere all'utente. Se il contesto dialogico e l'interpretazione del parlato spontaneo rendono la progettazione di sistemi di dialogo un lavoro complesso e di continuo compromesso tra robustezza dei risultati ed elasticità, è proprio l'architettura del sistema che può risolvere le ambiguità dell'interazione e migliorare la *performance* degli altri moduli.

Il compromesso tra risultati ed elasticità è strettamente connesso con l'iniziativa concessa all'utente, che determina ciò che può dire in relazione agli enunciati del sistema: per questo, prima ancora di affrontare le diverse architetture, è necessario focalizzare brevemente sul concetto di iniziativa, sia da un punto di vista teorico, sia da un punto di vista pratico.

2.4.1 Gestione dell'iniziativa

2.4.1.1 L'iniziativa da un punto di vista teorico

Il concetto di iniziativa, pur essendo nato nell'ambito della linguistica computazionale e dei sistemi automatici di dialogo, è affine al concetto di preferenza nel completamento di una sequenza complementare (Schegloff – Sacks 1973). In una conversazione, il turno di uno dei partecipanti può condizionare con più o meno forza il turno del parlante successivo, nel quale l'interlocutore selezionato deve innanzitutto soddisfare le aspettative del turno antecedente, sulla base del principio di cooperazione. Un parlante detiene l'iniziativa in un dato momento del dialogo quando indirizza e determina, parzialmente, il turno dell'interlocutore successivo (Burke 1994) e localmente la risoluzione del compito (Novick – Sutton 1997); il passaggio dell'iniziativa avviene quando il parlante, nel suo turno, non si limita a rispondere all'interlocutore, e quindi a soddisfarne le aspettative, ma presenta a sua volta un intento che obbliga l'interlocutore a soddisfarlo (Walker – Whittaker, 1990).

Sulla base di questa meccanica di gestione delle aspettative, Walker e

Whittaker, i primi a cercare di codificare l'iniziativa, hanno classificato alcuni tipi di enunciati basandosi sulle regole di controllo che questi esercitano sul parlante successivo:

- asserzione: iniziativa del parlante, a meno che non sia la risposta a una domanda;
- comando: iniziativa del parlante;
- domanda: iniziativa del parlante, a meno che non sia la risposta a una domanda o a un comando;
- risposta: iniziativa dell'ascoltatore (ovvero, colui che nel turno precedente ha posto la domanda).

Successivamente, sempre nell'ambito della linguistica computazionale, sono stati elaborati diversi modelli teorici con diverse definizioni di iniziativa (Cohen et al. 1999), che oscillano tra due poli principali: da una parte l'iniziativa vista come una meccanica relativa al flusso della conversazione e al suo controllo locale, dall'altra l'iniziativa vista come una meccanica che definisce chi sta guidando la risoluzione del compito, presentando obiettivi e piani d'azione all'altro interlocutore.

Secondo i modelli basati sul controllo della conversazione (Walker – Whittaker 1990), detiene l'iniziativa chi gestisce la direzione del flusso della conversazione: il passaggio d'iniziativa avviene quando un partecipante cambia la direzione della conversazione. Le diverse direzioni che può avere un turno rispetto al flusso della conversazione sono cinque: andare avanti (continuare nella direzione della conversazione, rispondendo alle domande o dando feedback di comprensione: in questo caso non avviene alcun passaggio d'iniziativa), cambiare la direzione (sospendendo il piano attuale e proponendo un sotto-piano o un piano rispetto al precedente: in questo caso l'iniziativa passa al parlante che ha proposto il cambio), sospendere la conversazione (chi detiene l'iniziativa può determinare la pausa della conversazione: in questo caso l'iniziativa rimane a chi ha sospeso la conversazione), chiuderla (se il parlante non ha più contributi, l'ascoltatore può prendere l'iniziativa o altrimenti chiudere la conversazione), interromperla (in tal caso, l'ascoltatore che ha interrotto la conversazione prende l'iniziativa). Al contrario, modelli più focalizzati

sulla risoluzione di un problema, vedono nell'iniziativa la capacità che hanno gli interlocutori di determinare la direzione dell'esecuzione di un compito: o proponendo soluzioni (Chu-Carroll – Brown 1997) o ponendo obiettivi da raggiungere.

2.4.1.2 L'iniziativa da un punto di vista applicativo

I modelli elaborati derivano direttamente dalle ricerche sull'iniziativa mista: nei dialoghi ordinari tra persone è normale che l'iniziativa (o il controllo della conversazione) slitti di parlante in parlante durante l'interazione (Walker & Whittaker, 1990), determinando di volta in volta chi, con il proprio turno, gestisce anche (in parte) il turno successivo e, localmente, la risoluzione del compito. In un sistema di dialogo ciò che il parlante può dire nel proprio turno è determinato dalla gestione dell'iniziativa prevista. Il sistema può essere a iniziativa singola, come nelle architetture a stati finiti (cfr. 2.4.2), e quindi decidere rigidamente cosa il parlante può dire, senza che questi abbia la possibilità di dare maggiori informazioni o cambiare la direzione della conversazione. Se l'iniziativa è solo in mano all'utente, allora si parla di *user-initiative system*: questo tipo di gestione è usata, più che nei sistemi di dialogo, nelle applicazioni per interrogare database. Successivamente ai primi sistemi a stati finiti, l'attenzione si è spostata sull'iniziativa mista, per permettere all'utente di gestire, anche se parzialmente, il proprio turno e quindi il flusso della conversazione. Tuttavia, una gestione a iniziativa mista “completa”, come quella descritta da Walker – Whittaker 1990, è ancora difficile da raggiungere; di conseguenza, si tende a usare il termine iniziativa mista anche per sistemi di dialogo che consentono solo parzialmente all'utente di prendere l'iniziativa (Cassell 2009). In genere, il termine è usato per descrivere quei sistemi di dialogo che combinano una domanda aperta iniziale non restrittiva, del tipo “How may I help you?”, usata per identificare l'argomento della conversazione, con successive domande dirette restrittive. Da un punto di vista applicativo, infatti, la gestione dell'iniziativa prevista determina di fatto quello che un utente può dire in un dato turno e la libertà che ha di condizionare la configurazione dei turni. Nei sistemi a iniziativa mista, generalmente, l'iniziativa è determinata dalle regole di comprensione

dei moduli di ASR e NLU: le domande del sistema sono scritte appositamente per instradare o meno il parlante sulle variabili da fornire al sistema, sulla base delle sue capacità di comprensione. La combinazione tra tipo di grammatiche di comprensione (restrittive o non restrittive) e tipo di domande del sistema, ha portato Singh et al. 2002 a proporre una classificazione per definire l'iniziativa di un sistema basata solo su parametri applicativi (Tab. 1):

	Tipo di domanda	
Grammatica	Aperta	Diretta
Restrittiva	***	Iniziativa del sistema
Non restrittiva	Iniziativa dell'utente	Iniziativa mista

Tab. 1: La gestione dell'iniziativa in relazione al modulo di comprensione e al dialogue manager

Se l'iniziativa non è ancora ben definita dal punto di vista teorico, trova la sua principale ragion d'essere proprio da un punto di vista applicativo, risultando strettamente legata alle grammatiche di comprensione del sistema e all'architettura di questo. Difatti, nei sistemi a iniziativa singola, a seconda della domanda del sistema, la grammatica è ristretta unicamente ai parametri richiesti relativamente allo stato del dialogo (ad es., “Da che città vuoi partire?” può avere come risposta unicamente una città); se il sistema è a iniziativa dell'utente, la domanda è aperta, perciò il sistema non può imporre restrizioni nella grammatica di comprensione. Un sistema a iniziativa mista è generalmente caratterizzato da domande dirette, pur prevedendo la possibilità per il parlante di cambiare la direzione della conversazione, o fornendo più informazioni di quelle che sono state esplicitamente richieste (che possono essere rilevanti o meno per il compito), o correggendo direttamente il sistema in qualsiasi momento dell'interazione: anche in questo caso la grammatica non può essere legata al singolo stato del dialogo, ma deve prevedere una più larga gamma di azioni conversazionali che il parlante può svolgere.

2.4.2 Architettura a stati finiti

L'architettura più semplice di gestione del dialogo è quella a stati finiti,

ampiamente usata nei centralini telefonici; deriva infatti dalle interfacce DTMF (*Dual Tone Multi Frequency*), che costringono l'utente a selezionare le opzioni proposte tramite il tastierino numerico del telefono, e dalle interfacce vocali, con cui l'utente può unicamente pronunciare le parole proposte dal sistema. In questa architettura è previsto che il sistema sia l'unico a gestire l'iniziativa: dirige infatti sia l'andamento della conversazione sia la risoluzione del compito, richiedendo all'utente, di volta in volta, le informazioni necessarie, con grammatiche di comprensione ristrette unicamente alla richiesta del sistema. Ad esempio, in un ipotetico servizio di prenotazione di biglietti del treno, può essere chiesta in successione, volta per volta, la città di partenza, la città d'arrivo, la data, fino alla richiesta di conferma finale e alla prenotazione del biglietto.

Essendo l'interazione gestita unicamente dal sistema, il parlante non può controllare la direzione della conversazione, come normalmente avviene nelle conversazioni tra parlanti; l'unico spazio che l'utente ha a disposizione per intervenire sul flusso della conversazione è concesso tramite comandi universali, come “aiuto” e “menu”, che permettono all'utente di riavviare la conversazione o di ricevere istruzioni su come usare il sistema in qualsiasi momento dell'interazione.

Un'architettura del genere, chiamata anche *command-and-control*, è vantaggiosa per la facilità di comprensione e per la robustezza dei risultati: il riconoscimento vocale e la comprensione automatica sono semplificati dalla sicurezza del sistema relativamente a ciò a cui si sta riferendo l'utente. Tuttavia, un dialogue manager così concepito è adatto principalmente a compiti semplici, in cui sono previsti pochi turni; l'adozione della stessa architettura per compiti più complessi, porta a dover prevedere un numero maggiore di turni necessari per completare il compito, dando luogo a interazioni molto lunghe e potenzialmente frustranti per l'utente, e annullando così i vantaggi che possono derivare dall'uso di un'interfaccia vocale rispetto ad altri tipi di interfacce.

2.4.3 Architettura *frame-based*

Un'architettura più elastica e che lascia più spazio all'utente, consentendo

l'iniziativa mista, è quella *frame-based* (già vista in parte in 2.3.2). Con questa architettura, il dominio e il relativo compito da svolgere sono organizzati secondo una suddivisione in *slot* che il parlante deve “riempire” nel proprio turno. Tale operazione può avvenire in un numero di turni minore rispetto all'architettura a stati finiti: la grammatica di comprensione consente infatti al parlante di fornire più dati di quelli esplicitamente richiesti dal sistema. La grammatica di comprensione presentata in 2.3.2, caratteristica di questa architettura, consente infatti di gestire risposte con più informazioni:

Sistema: Da che città vuoi partire?

Utente: Voglio partire da Roma il 27 aprile e andare a Milano

Nel caso in cui, durante l'interazione, uno degli slot necessari per lo svolgimento dell'attività risulti scoperto, il sistema chiede esplicitamente un valore per quel determinato slot.

Sistema: Da che città vuoi partire?

Utente: Voglio partire da Roma il 27 aprile e andare a Milano

Sistema: Hai preferenze per l'orario?

Utente: Sì, la mattina

Come abbiamo visto relativamente al modulo di comprensione del linguaggio, il frame può essere inquadrato come una classe semantica con gli slot da riempire relativamente a una determinata attività da svolgere: ad esempio, per avere delle informazioni su una determinata tratta ferroviaria devono essere riempiti gli slot relativi alla data, all'ora, la città di partenza e a quella di arrivo; ulteriori informazioni possono essere collocate in un altro frame, che avrà slot relativi alla classe sulla quale si vuole viaggiare, al tipo di treno, ecc. Quando un sistema deve svolgere compiti più complessi su domini più estesi, una tale suddivisione del lessico può portare a conflitti tra classi: uno slot può infatti riguardare più classi semantiche (ad esempio la data e l'ora possono riguardare sia il viaggio di andata che quello di

ritorno). Il sistema deve essere quindi in grado di collocare i valori estratti dall'enunciato del parlante nel frame giusto: per questo i sistemi sono spesso dotati di regole di controllo del dialogo (Seneff – Polifroni 2000) che permettono di disambiguare l'enunciato del parlante (ad esempio, una volta identificato che si tratta del viaggio di andata, i valori relativi alla data e all'ora andranno a riempire lo slot relativo a quel frame). Modelli di interazione alternativi a quello a iniziativa mista possono essere previsti per determinate classi di utenti o in alcuni punti dell'interazione: ad esempio, le grammatiche di comprensione non sono ristrette al singolo stato del dialogo, tuttavia, in determinati punti cruciali dell'interazione (come la conferma di una transazione o di una prenotazione) possono essere previste delle grammatiche più restrittive, che obbligano il parlante a confermare o a rifiutare. Anche se la gestione dell'iniziativa è mista, l'iniziativa del sistema può essere inoltre prevista per quei parlanti che non cercano il controllo dell'interazione e preferiscono lasciarsi guidare dal sistema nello svolgimento dell'attività (cfr. 3.4). L'iniziativa singola può inoltre essere prevista dopo un certo numero di errori: in tal caso, il sistema, oltre a prendere il controllo dell'interazione, suggerisce al parlante termini che può usare.

2.4.4 Gestione delle incomprensioni¹²

Nelle conversazioni ordinarie tra persone è normale che molte delle informazioni non comprese direttamente dal ricevente siano recuperate tramite la conoscenza del contesto; il dialogo uomo-macchina implica invece numerosi passaggi problematici: errori del riconoscitore, rumori, errori di interpretazione pragmatica, errori del parlante stesso. Anche i sistemi più semplici, a stati finiti, con un lessico molto limitato non raggiungono il 100% di tasso di successo: chiunque abbia provato un'interfaccia vocale sa quanto siano normali e frequenti gli errori del sistema. Per questo la gestione delle incomprensioni è un aspetto chiave di un sistema di dialogo, che incide sui risultati e sulla capacità di un sistema.

Come già visto, il riconoscitore vocale assegna un valore di *confidence* a ogni

¹² In questo paragrafo vengono riassunte brevemente le strategie di risoluzione degli errori, per una descrizione più esaustiva delle tipologie di errore e delle strategie di risoluzione si rimanda a 6.4.

input del parlante (cfr. 2.3.1), a cui corrisponde l'affidabilità delle informazioni acquisite dal sistema: questa è ottenuta dal riconoscimento acustico e da previsioni basate sulla statistica a N-grammi, oltre che naturalmente dallo stato del dialogo. I valori di confidence vengono usati per fare da filtro tra gli input da acquisire e gli input che invece necessitano una conferma o una riformulazione. A fronte di un enunciato ambiguo, la gestione degli errori procede principalmente tramite due tipi di richieste: le richieste di ripetizione e le richieste di conferma.

Le richieste di conferma vengono selezionate quando la confidence è intermedia tra la comprensione e l'incomprensione: se il sistema non è sicuro dei valori estratti dall'enunciato del parlante, chiede all'utente di confermare l'informazione acquisita, spesso passando a una modalità a iniziativa del sistema, che costringe il parlante a usare solo “sì” o “no”.

Sistema: Dove vuole arrivare?

Utente: Voglio arrivare a Milano

Sistema: Vuole andare a Merano? Dica sì o no

Utente: No

Sistema: Dove vuole arrivare?

In architetture più elastiche, l'iniziativa mista viene lasciata anche durante la gestione della conferma da parte dell'utente: in tal caso il sistema è in grado di interpretare, oltre alla disconferma o alla conferma, ulteriori informazioni, che possono essere la correzione diretta o nuove informazioni che il parlante fornisce nel suo turno oltre alla conferma:

Sistema: Vuole andare a Merano?

Utente: No, voglio andare a Milano

Oltre alle richieste esplicite, la conferma può essere gestita tramite richieste implicite, in cui si chiede conferma dei dati acquisiti indirettamente, presentandoli nel turno successivo in cui il sistema pone un'altra domanda, mostrando così la

variabile acquisita:

Utente: Voglio andare a Milano

Sistema: Quando vuoi partire per Merano?

Con le richieste implicite, le potenziali ambiguità possono essere gestite più agilmente e con meno turni, in maniera più naturale e con il vantaggio per il parlante umano di avere continui feedback su ciò che ha detto (Danieli – Gerbino 1995); d'altra parte, come abbiamo già visto in numerosi punti di questo capitolo, a maggiore elasticità corrisponde generalmente una minore robustezza dei risultati.

Quando la confidence è talmente bassa da rendere scarsamente produttiva una richiesta di conferma per un'informazione molto probabilmente sbagliata, il sistema può comunicare all'utente di non aver capito e chiedere di riformulare la frase. Anche in questo caso le strategie possono essere diverse: ad esempio, a scapito della velocità e della naturalezza, ma in favore di una maggiore robustezza dei risultati, piuttosto che richiedere la ripetizione dell'enunciato, può essere più funzionale che nella richiesta di ripetizione vengano incluse istruzioni su ciò che il parlante può dire (Weinschenk – Barker 2000):

Sistema: Quando vuoi partire?

Utente: Il 28 aprile di mattina

Sistema: Non ho capito. Per favore dimmi la data e l'ora.

Questa strategia viene generalmente usata dopo che l'errore non è stato risolto tramite una semplice richiesta di ripetizione o con il metodo del *rapid reprompting*, con cui il sistema “finge” di non aver sentito, tramite un turno veloce e breve “Come?” “Cos'hai detto?”, ecc. (Coehn et al. 2004).

2.4.5 La comprensione pragmatica

Le architetture viste finora si basano su una comprensione lessicale

dell'enunciato del parlante: ciò le rende applicabili principalmente su sistemi che operano su domini limitati e facilmente formalizzabili in classi semantiche. Tuttavia, come sappiamo dal dialogo uomo-uomo, l'interpretazione di un enunciato non si basa solo sul contenuto proposizionale, ma anche sul contesto, locale e globale. Per quanto i sistemi di dialogo commerciali siano ancora lontani dall'essere provvisti di una capacità di comprensione pragmatica, alcune architetture sperimentali permettono un'interpretazione dell'enunciato basata anche su ciò che non è stato esplicitamente detto dal parlante e sul suo ruolo pragmatico. Tali modelli si basano sui *dialogue acts* ('atti di dialogo', a volte chiamati *conversational acts*, *dialogue moves* o *conversational moves*), un modello secondo cui gli enunciati, oltre al contenuto proposizionale, contengono delle indicazioni che incidono sullo svolgimento dell'interazione e che vanno soddisfatte.

Una di queste, senz'altro la più importante, è l'architettura *information-state* (Traum 2000, Traum – Larsson 2003). A differenza dei modelli precedenti, più rigidi, l'architettura *information-state* non si basa unicamente sulle informazioni esplicite fornite, ma è in grado di fare ragionamenti sulle intenzioni dell'utente: in base a ciò che l'utente ha detto esplicitamente (e viene quindi condiviso), in base alle conoscenze del dominio e del compito, in base al contesto del discorso e in base a un aggiornamento continuo delle mosse da fare conseguentemente agli atti del parlante.

Questa architettura è composta, oltre che dai moduli già visti, da altre componenti che si occupano della comprensione pragmatica e del comportamento conversazionale. Al centro vi è il modulo *information-state*, che, oltre a interfacciarsi, come nell'architettura normale, con il dominio e i compiti da svolgere, gestisce i dati provenienti dal modulo di comprensione degli atti di dialogo e decide quali atti di dialogo generare; per fare questo, il modulo *information-state* necessita di un altro modulo che aggiorna il contesto del dialogo sulla base degli atti di dialogo eseguiti dal parlante: ad esempio, a un'asserzione corrisponde, oltre che l'acquisizione del contenuto semantico, un aggiornamento del contesto del dialogo, che comporta la produzione di un feedback con cui il sistema mostra di aver compreso.

2.4.5.1 Dialogue acts

Per quanto il modello sia apparentemente semplice, le difficoltà maggiori nella definizione e classificazione dei dialogue acts. Un atto di dialogo è la combinazione tra la funzione comunicativa e il contenuto semantico di un enunciato: concettualmente equivalenti alla forza illocutoria e al contenuto proposizionale di un enunciato, su cui si fonda la teoria degli atti linguistici. Gli atti di dialogo non hanno uno statuto teorico ben definito, ma risultano utili per descrivere il significato di un enunciato e come questo cambi la rappresentazione del contesto (Bunt 2000).

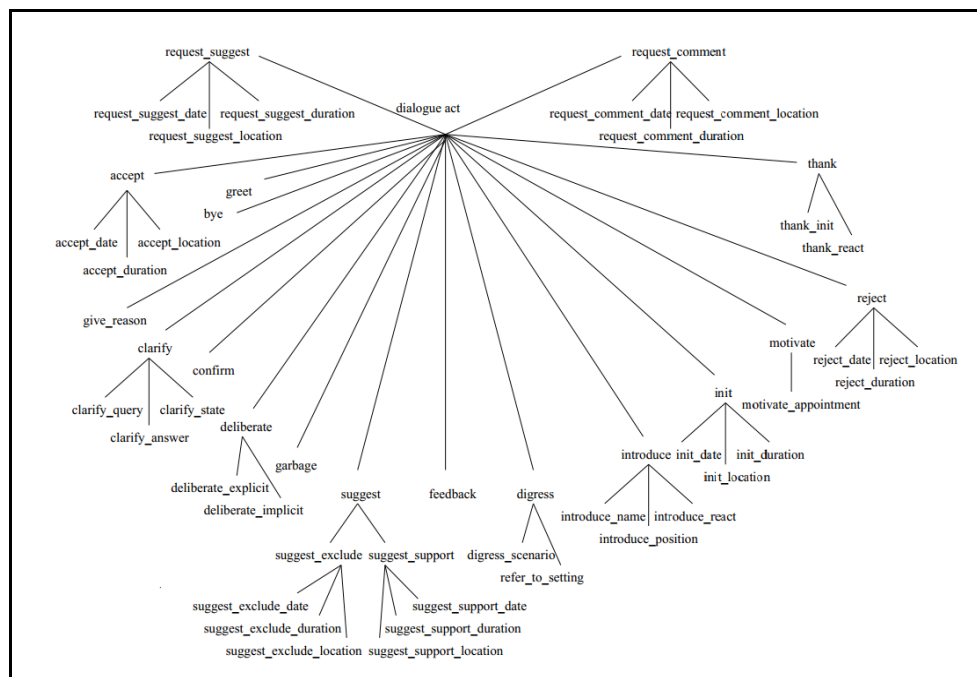


Fig. 2: Classificazione degli atti di dialogo nel progetto Verbmobil (da Jekat et al. 1995)

Le classificazioni degli atti di dialogo possono essere legate a un dominio specifico o a uno specifico sistema; altre classificazioni ambiscono invece ad essere universali e slegate da contesti particolari. Le classificazioni specifiche sono più accurate ma difficili da trasportare da un dominio (o un sistema all'altro): quella elaborata per il progetto Verbmobil (fig. 2, Jekat et al. 1995), ad esempio, prevede ben 43 tipi di atti diversi, alcuni generali (come ad esempio CONFIRM o FEEDBACK) e altri specificamente relativi al sistema (come ad esempio

SUGGEST_SUPPORT_DATE o INTRODUCE_NAME). Un esempio, in questo caso italiano, di classificazione legata a un dominio, è il Pr.A.T.I.D. (Savy 2010), elaborata per etichettare dialoghi elicitati con la tecnica del map-task.

Modelli più generali come il DAMSL sono invece stati elaborati per coprire più domini (Allen – Core 1997). Questa classificazione prevede alcune categorie non legate direttamente alle funzioni comunicative: come lo stato comunicativo (ovvero se l'enunciato è interpretabile o meno), il livello informativo (se l'enunciato riguarda direttamente l'esecuzione del compito, il parlare a proposito dell'esecuzione del compito, o la funzione fatica di mantenimento del canale). Le funzioni comunicative sono invece suddivise in due macro-categorie:

³⁵₁₇ *forward-looking dialogue acts*: riguardano l'effetto che può avere l'enunciato sulla struttura futura dell'interazione, come ad esempio asserzioni, richieste di informazioni, comandi, suggerimenti, promesse, ecc.

³⁵₁₇ *backward-looking dialogue acts*: riguardano le relazioni tra l'enunciato e quanto detto precedentemente, come l'accordo o il disaccordo, la comprensione, le risposte a una richiesta e la loro relazione con il contenuto informativo antecedente.

Un'altra classificazione, proposta da Traum – Hinkelman 1992, include gli atti linguistici base (come richiesta, comando, asserzione, ecc.) e gli atti legati al *grounding*¹³ in una classificazione di atti legati alla gestione della conversazione, che prevede, in aggiunta, altre due classi: gli atti legati alla gestione del turno e gli atti legati alle relazioni di coerenza con l'antecedente (chiarificazioni, rielaborazioni, ecc.).

I ricercatori, oltre a scontrarsi con la complessità strutturale di una classificazione degli atti di dialogo (simile alla grande varietà di classificazioni proposte per gli atti linguistici in generale), trovano ulteriori difficoltà nel renderli individuabili e interpretabili dal sistema. Infatti, se l'uso degli atti di dialogo cerca di

¹³ Si intendono, per atti legati al *grounding*, tutti i segnali linguistici che usano gli interlocutori per mostrare ciò che viene acquisito e ciò che non viene acquisito durante la conversazione (da *common ground*, cfr. Stalnaker 1978; Clark – Schaefer 1989)

superare una comprensione basata sul semplice contenuto semantico dell'enunciato, individuato principalmente dal lessico usato, già formalizzato in frame, è necessario trovare altre caratteristiche osservabili dal sistema che permettano a questo di capire quale atto di dialogo sia stato usato dal parlante. Anche in una lingua come l'inglese, che impone rigide regole sintattiche (come l'inversione tra soggetto e ausiliare per le domande dirette e la sequenza soggetto e ausiliare per le asserzioni) è frequente, nel contesto dialogico, che vengano usati atti linguistici indiretti, come ad esempio una dichiarativa usata per una domanda diretta (Carletta et al. 1997).

Per interpretare gli atti di dialogo viene usata una microgrammatica, addestrata su un *corpus* etichettato a mano, da cui si estraggono caratteristiche relative a ogni atto di dialogo, basate sulla forma dell'enunciato e sul contesto locale: parole specifiche e sintassi (individuate da grammatiche a N-grammi), prosodia, struttura conversazionale. Ulteriori grammatiche possono essere scritte per specifici atti di dialogo, come ad esempio le correzioni del parlante quando il sistema fa un errore. Per migliorare l'individuazione delle correzioni occorre quindi studiare le strategie di riformulazione dei parlanti (ad esempio, Swerts et al. 2000 riportano che sono più frequenti le ripetizioni della frase e le ripetizioni con omissione di elementi considerati superflui, rispetto alla parafrasi). Il modulo che individua le autocorrezioni è generalmente basato su indicatori lessicali (in inglese forme come “no”, “correction”, “I don't”), caratteristiche prosodiche (come lunghezza delle pause o lunghezza delle sillabe, Levow et al. 1998), lunghezza delle frasi, oltre ai valori di confidence che stabilisce l'ASR.

2.5 Siri

Il sistema di dialogo attualmente più diffuso e famoso è senz'altro SIRI, applicazione di serie dello smartphone iPhone. Il progetto è nato nel 2007, grazie a un gruppo di ricercatori provenienti dai laboratori dello SRI (Stanford Research Institute) che hanno fondato l'azienda Siri con lo scopo di sviluppare un assistente vocale da poter distribuire come applicazione per smartphone. Nel 2010

l'applicazione è stata lanciata sull'Apple Store e, dato il grande successo, è stata acquistata da Apple e lanciata come applicazione di serie con l'iPhone 4S. Essendo un prodotto commerciale non è purtroppo disponibile letteratura scientifica sull'architettura del sistema.

Il sistema può essere attivato con tre modalità, nessuna di queste verbale: premendo un tasto per un secondo o portando il telefono all'orecchio come se si stesse telefonando. L'attivazione è comunicata da un segnale non verbale (un doppio bip) e da una domanda (solo scritta) “Come posso aiutarti?”. Dopo che l'utente ha parlato (successivamente al doppio bip di attivazione), Siri dà un feedback di avvenuta ricezione (anche in questo caso non verbale) ed elabora l'input.¹⁴

Siri può lavorare su numerosi domini, alcuni integrati e "di base" dell'iPhone 4S, altri via applicazioni esterne (come Yahoo! per lo sport e la borsa, o Yelp per ristoranti e negozi). Se l'input iniziale soddisfa tutti i requisiti dell'azione che si intende svolgere, l'azione viene eseguita al primo turno, altrimenti, individuato il dominio su cui svolgere il compito, vengono richieste le informazioni necessarie per l'interazione tramite domande dirette. Questo fa sì che i dialoghi siano differenti a seconda del dominio su cui si sta svolgendo l'interazione. Possiamo comunque distinguere tra domini in cui l'interazione è unicamente *command-and-control* (cfr. 2.4.2) e dialoghi effettivi, con la creazione di un background dialogico condiviso.

I dialoghi *command-and-control* riguardano principalmente le applicazioni “esterne” all'iPhone: è infatti possibile avviare qualsiasi applicazione con Siri. Una volta aperta l'applicazione, senza alcun feedback vocale, il dialogo si chiude. Altre applicazioni (come la ricerca di ristoranti o le informazioni sportive), pur con un approccio *command-and-control*, restituiscono un feedback con le informazioni richieste, in formato parziale, mentre il contenuto completo è visibile graficamente. Tuttavia non si tratta di veri e propri dialoghi: Siri infatti non tiene memoria dell'interazione, per cui ogni richiesta di informazioni è a sé stante.

Le applicazioni integrate (come il telefono, la rubrica, i messaggi, ecc.) consentono, a differenza delle altre, un dialogo più complesso, che tiene memoria della conversazione sostenuta. Ad esempio, per il telefono sono previste richieste di

¹⁴ Il riconoscimento vocale (sviluppato dall'azienda Nuance, anche se tale informazione non è stata mai confermato in via ufficiale da entrambe le parti in causa) avviene in remoto.

disambiguazione nel caso il nome a cui telefonare corrisponda a più numeri, richieste di conferma una volta selezionato il numero, ecc.

Il sistema è a iniziativa mista totale completa che prescinde dal tipo di azione che si sta svolgendo: quando si sta svolgendo un'operazione che prevede più di un turno d'interazione, per l'utente è possibile passare a un'altra applicazione con un solo enunciato (tranne quando è attivo il sistema di dettatura). Tuttavia, l'architettura estremamente flessibile è controbilanciata da grammatiche eccessivamente restrittive in alcune fasi del dialogo. Infatti, la grammatica del modulo di comprensione prevede per alcuni turni del parlante un approccio *command-and-control*. Ad esempio, una volta dettato un messaggio, l'output successivo del sistema "Per continuare puoi inviare, annullare, modificare l'oggetto, modificare il messaggio o aggiungere qualcosa." prevede per l'utente la sola possibilità di selezionare tra i termini proposti dal sistema, senza peraltro informare esplicitamente l'utente di tale costrizione.

Le incomprensioni sono gestite diversamente a seconda della tipologia a cui appartengono: assenza di segnale audio, incomprensione o comprensione di input fuori dominio (o ritenuti tali da Siri). Se Siri non individua alcun input, si disattiva dando unicamente un segnale di feedback di disattivazione (un doppio bip con una frequenza più bassa); se invece decodifica un segnale vocale che non riesce a ricondurre al lessico, si limita a segnalare per via vocale e scritta di non aver capito ("Non ho capito bene", "Spiacente, non capisco", ecc.), chiudendo il canale di comunicazione. Nel caso in cui l'input viene compreso ma non associato a uno dei servizi coperti da Siri, viene riportato l'input trascritto, chiedendo se si vuole cercare tale frase su internet (alternativa non prevista con la modalità cuffie e modalità telefono, in cui viene semplicemente riportato l'input compreso con la segnalazione dell'incomprensione).

3. Le interazioni verbali tra parlanti umani e sistemi di dialogo automatici

3.1 Introduzione

La prima interfaccia per l'interazione con un sistema artificiale tramite l'uso del linguaggio naturale¹⁵, ELIZA, fu presentata su un articolo dal titolo “A Computer Program For the Study of Natural Language Communication Between Man And Machine” (Weizenbaum 1966). Il prototipo serviva a mostrare un modello per l'uso di un software tramite comandi in linguaggio naturale, ed era concepito allo stesso tempo per studiare tali comandi e le possibili implicazioni per la programmazione di interfacce simili (cfr. *infra*). Il prototipo era stato concepito come strumento per l'elicitazione di input, allo scopo di studiare le caratteristiche del linguaggio naturale da tenere in considerazione per sviluppare nuove interfacce meno costrittive di quelle a comandi (CLI, *Command Line Interfaces*), che, al contrario, richiedevano che l'utente conoscesse il linguaggio del sistema operativo. Il modello di interazione si basava sul metodo della psicoterapia rogersiana, che, allo scopo di far esprimere il paziente condizionandolo il meno possibile, prevede una serie di domande aperte basate sugli elementi già presenti nella risposta del paziente. Un sistema con lo stesso meccanismo di interazione poteva, in tal modo, non essere provvisto di alcuna conoscenza esterna all'interazione stessa, e acquisire i dati dell'utente unicamente in funzione della domanda successiva:

U: Men are all alike
E: In what way
U: They're always bugging us about something or other
E: Can you think of a specific example
U: Well, my boyfriend made me come here
E: Your boyfriend made you come here
U: He says I'm depressed much of the time
E: I am sorry to hear you are depressed

¹⁵ Il concetto di “linguaggio naturale”, centrale in questo capitolo e, come già visto (cap. 1), di grande uso nella linguistica computazionale, verrà discusso nel paragrafo 3.4.

U: It's true, I am unhappy
 E: Do you think coming here will help you not to be unhappy?
 U: I need some help, that much seems certain
 E: What would it mean to you if you got some help
 U: Perhaps I could learn to get along with my mother
 E: Tell me more about your family
 U: My mother takes care of me
 E: Who else in your family takes care of you
 U: My father

Nell'esempio le risposte del sistema sono coerenti con gli input del parlante e il dialogo può sembrare quello tra due esseri umani. Lo stesso Weizenbaum notava come non fosse eccessivamente difficile simulare una reale interazione semplicemente riprendendo un elemento dell'input e mettendolo al centro della successiva richiesta.

Prima ancora dello sviluppo delle interfacce grafiche (GUI, *Graphic User Interfaces*), e poi parallelamente alla loro prima diffusione, il modello per lo sviluppo di interfacce in linguaggio naturale è stato quello del parlante umano e delle sue competenze linguistiche: un agente artificiale che eliminasse la necessità di apprendere altri codici per interagire con computer e sistemi artificiali. Il sistema perfetto avrebbe dovuto consentire al parlante «to forget that he is questioning a machine» (Gal 1988), tramite interazioni simili a quelle che intratteniamo quotidianamente con altri parlanti umani. Lo stesso Test di Turing pone come parametro di valutazione dell'intelligenza artificiale la similarità tra le competenze linguistiche di un essere umano e quelle di un sistema (dando per scontata una sostanziale equivalenza tra “comprendere” e “rispondere in maniera conversazionalmente accettabile”).¹⁶ Tuttavia, le capacità dialogiche di un sistema devono innanzitutto servire a svolgere le funzioni di interfaccia, prima ancora che di interlocutore: per descrivere il dialogo uomo-macchina è necessario focalizzare innanzitutto sulla situazione comunicativa generalmente coinvolta.

¹⁶ A tal proposito è interessante lo studio fatto da Saygin – Cicekli 2002 sui risultati del premio Loebner: nell'esperimento vengono poste in relazione la violazione delle massime della conversazione di Grice e la percezione da parte dei giudici di star interagendo con un essere umano (ad esempio, la violazione della massima di cortesia viene percepita come manifestazione di emozioni, e quindi aumenta la verosimiglianza di star interagendo con un essere umano).

3.2 Situazioni comunicative e tipi di interazione tra umani e interfacce vocali

Lo sviluppo di altri tipi di interfacce, alternativi alle CLI (come quelle GUI), e una maggiore attenzione per l'usabilità e per le situazioni reali in cui gli utenti interagiscono con sistemi artificiali (soprattutto a partire da Norman 1988 e dalla diffusione dell'*user-centered design*, la progettazione incentrata sull'utente) rendono necessario confrontarsi con i tratti strutturali e con le particolarità che caratterizzano il dialogo uomo-macchina, più funzionale per la progettazione di un sistema rispetto alle caratteristiche del dialogo uomo-uomo.

Le capacità conversazionali e linguistiche di un sistema artificiale non devono coprire tutte le interazioni possibili in cui quotidianamente sono coinvolti gli esseri umani: l'interazione con una macchina è infatti una situazione comunicativa particolare rispetto alle situazioni comunicative quotidiane, sia per situazione extralinguistica sia per svolgimento effettivo del dialogo. Un sistema di dialogo serve innanzitutto a svolgere una determinata attività: questo riduce notevolmente le capacità necessarie per operare come interfaccia.

Le sostanziali differenze tra dialoghi uomo-uomo e dialoghi uomo-macchina, che attualmente costituiscono il punto di partenza per lo sviluppo di qualsiasi interfaccia vocale, si collocano su vari livelli comunicativi: nel capitolo verranno prima individuate le caratteristiche peculiari del contesto d'interazione uomo-macchina, a partire dal *setting* e dai tipi di dialogo; successivamente verranno espone le caratteristiche linguistiche del dialogo uomo-macchina individuate negli studi empirici condotti, che lo caratterizzano generalmente come “semplificato”; infine verrà approfondito il legame tra il comportamento linguistico del parlante e la sua rappresentazione della situazione e del sistema, rappresentazione che oscilla tra due poli: il sistema in quanto interfaccia e il sistema in quanto interlocutore.

3.2.1 La situazione extralinguistica

Per situazione extralinguistica intendiamo i «vari aspetti della situazione interazionale: tempo, spazio, mezzo fisico di trasmissione, partecipanti» (Bazzanella 2005: 2). La presenza di un sistema automatico nel ruolo di uno dei due interlocutori è certamente il tratto strutturale più distintivo del dialogo uomo-macchina, che

definisce, inoltre, la maggiore o minore rilevanza nell'interazione di altri tratti del *setting*. Nell'analisi del dialogo uomo-macchina, alcune caratteristiche, determinate dell'architettura del sistema e delle capacità dei vari moduli di comprensione, possono infatti essere considerate *a priori* e non sono negoziabili durante l'interazione. Le proprietà del sistema, oltre a incidere direttamente sull'interazione dal punto di vista linguistico (dalla comprensione degli input alla produzione degli output), determinano la rilevanza o meno di altri tratti che caratterizzano la situazione, come la condivisione spaziale e temporale. La reale copresenza fisica dei due interlocutori, infatti, può non essere rilevante se non viene sfruttata dalle capacità di comprensione del sistema (ad esempio, la possibilità di sovrapporsi al turno del sistema o la rilevanza della gestualità), e conseguentemente non essere rilevante per l'utente. Naturalmente la presenza o l'assenza di tali caratteristiche non implica automaticamente la presenza o l'assenza di determinati fenomeni, ma ne può condizionare la presenza, una volta che l'utente ha appreso ciò che per il sistema è rilevante.

L'interazione avviene generalmente tra due soli interlocutori; tuttavia, va tenuta in considerazione la possibile presenza di altre persone non direttamente coinvolte nell'interazione, che può portare a casi di *offtalk* (Batliner et al. 2006) ossia parlato non rivolto alla macchina, o più in generale configurarsi come una *audience* (Duranti 1986) che può alterare parzialmente gli intenti comunicativi dell'interlocutore principale e il linguaggio usato verso il sistema.

L'interazione può avvenire in copresenza spaziale (ad esempio, con un'interfaccia vocale su smartphone) o senza (ad esempio, con un centralino telefonico automatico). Nel caso di copresenza spaziale può essere rilevante la presenza di un *avatar* umano (o non umano, infatti possono essere animali o personaggi di fantasia); nel caso di assenza dell'immagine dell'interlocutore, al contrario, può attenuarsi la rilevanza della copresenza. Inoltre, i sistemi di dialogo possono decodificare o meno input extralinguistici (in tal caso si parla di sistemi multimodali) come il movimento delle mani, della testa o degli occhi (importanti, ad esempio, per la gestione del turno). La condivisione spaziale è invece rilevante (sia per il sistema sia per il parlante) quando il sistema automatico deve essere in grado di interagire con l'ambiente circostante, come ad esempio quando costituisce l'interfaccia di un robot (Fischer 2007).

Allo stesso modo può essere più o meno rilevante la sincronia temporale,

soprattutto in relazione alla gestione del turno: in questo caso troviamo su lati opposti da una parte le interfacce vocali con una presa del turno *push-to-talk* (cfr. 6.3.3.3) o che emettono un segnale acustico per comunicare la cessione del turno, dall'altra quelle che consentono il *barge-in* (ovvero, la possibilità di interrompere il turno del sistema, cfr. 6.3.3.3), rendendo funzionale e rilevante la sovrapposizione. Nel primo caso l'interazione sarà più vicina alle interazioni asincrone, nel secondo, invece, presenterà fenomeni simili a quelle sincrone.

Infine, come per il dialogo uomo-uomo, il mezzo di trasmissione può essere scritto o vocale. Quando l'interazione avviene per via scritta il sistema viene comunemente denominato *chatbot* (o talvolta solo *bot*): trattandosi di un mezzo molto simile alla chat, vi si possono trovare i medesimi fenomeni della lingua delle chat (*typo*, espedienti tachigrafici, abbreviazioni, mancanza di programmazione sintattica e testuale). Tuttavia, a differenza delle interazioni mediate dalle chat tra utenti umani, nel caso dell'interazione con un chatbot non è consentito lo spezzettamento del turno su più linee di testo; il sistema, infatti, identifica ogni parte di testo inviata come un input.

Quando l'interazione avviene per via vocale può essere rilevante il tipo di microfono usato per l'acquisizione del segnale audio, che può essere quello di un telefono, oppure un radiomicrofono da portare al collo o un microfono direzionale: la posizione rispetto all'utente può infatti portare a comportamenti diversi del parlante, rispetto alle sue abitudini.

3.2.2 Tipo d'interazione

Anche il tipo di interazione è in parte determinato dalla presenza di un interlocutore non-umano e dalle specifiche che ne regolano il comportamento, come ad esempio le meccaniche di gestione del turno e dell'iniziativa; al contrario, altri tratti legati al tipo di interazione, come il compito e l'argomento (o dominio, per usare un termine più comune in linguistica computazionale), possono essere descritti indipendentemente dal tipo di sistema automatico presente nella conversazione.

Il genere di dialoghi che comunemente caratterizza l'interazione tra parlanti umani e sistemi automatici è stato definito *practical dialogue* da Allen et al. 2000. I dialoghi pratici possono essere distinti da quelli definiti, per comodità, “ordinari”¹⁷

¹⁷ Etichetta di comodo che ha valore in contrapposizione ai dialoghi pratici (cfr. Leech – Weisser 2005: 146).

per via del forte orientamento verso la risoluzione di un compito (la richiesta di informazioni, l'acquisto di un prodotto, ecc.) e per la delimitazione netta dei domini su cui vertono. Il livellamento sul compito e la limitazione del dominio portano, quindi, da una parte alla riduzione della variabilità linguistica e dall'altra a una macrostruttura *grosso modo* omogenea che prescinde dai domini:

«The practical dialogue hypothesis: the conversational competence required for practical dialogues, while still complex, is significantly simpler to achieve than general human conversational competence.» (Allen et al. 2000)

Un'utile schematizzazione pragmatica dei dialoghi pratici è stata proposta nell'ambito del progetto VERBMOBIL (Patzols et al. 1995, Alexandersson et al. 1997); ogni dialogo pratico è composto da cinque fasi: (1) saluti, (2) apertura del compito, (3) negoziazione, (4) chiusura del compito, (5) saluti. Escludendo i saluti¹⁸, gli altri stati della conversazione sono strettamente collegati al compito che l'utente deve svolgere.

L'apertura coincide con la coppia adiacente in cui viene individuato il compito: apertura che può riguardare la macchina (la quale può indirizzare l'utente sul compito che è in grado di svolgere), o riguardare l'utente (il quale, rispondendo a una domanda generica del tipo "Come posso aiutarti?", segnala, già nel primo turno, il compito che ha intenzione di svolgere, del tutto o in parte). Una volta aperto il compito inizia la fase di negoziazione, caratterizzata da richieste di completamento (volte a ricavare i dati mancanti dal parlante per lo svolgimento dell'attività), richieste di conferma e di ripetizione per la gestione degli errori; la negoziazione finisce con il completamento del task, segnalato dal sistema con un output di feedback sull'esecuzione di un'azione ("Ho aperto le finestre in cucina") o con la presentazione delle informazioni richieste ("Sono disponibili nove treni per Milano domattina").

Il tipo di interazione può inoltre essere parzialmente condizionato dal compito e dallo scopo della conversazione: la focalizzazione sul completamento di un'attività è la caratteristica principale dei dialoghi pratici. Gli attuali sistemi sono in grado di svolgere vari tipi di attività e compiti; in EAGLES 1998 vengono indicate tra le

¹⁸ I saluti possono essere rilevanti sia per caratterizzare un sistema come interlocutore (cfr. 3.4.3); allo stesso modo, l'eventuale risposta può essere utile per comprendere l'idea che il parlante ha della macchina e quindi prevedere possibili strategie, cfr. Fischer – Bateman 2006.

principali attività: negoziazione cooperativa (*cooperative negotiation*), estrazione di informazioni (*information extraction*), risoluzione di un problema (*problem solving*), istruzione, consulenza, ecc. A queste attività possono essere collegate alcuni compiti, alcuni dei quali sono ormai ritenuti caratteristici: gestione degli appuntamenti, pianificazione di un viaggio, richieste di informazione su aerei, pullman e treni, assistenza tecnica per specialisti, arredamento di una stanza, ecc.

Il dominio o l'argomento è anch'esso strettamente collegato al compito e all'attività da svolgere, soprattutto da un punto di vista lessicale: tuttavia, per quanto riguarda lo svolgimento dell'interazione, non sembrano esserci importanti condizionamenti dovuti al dominio, come suggeriscono Allen et al. 2000:

«The domain-independence hypothesis: within the genre of practical dialogue, the bulk of the complexity in language interpretation and dialogue management is independent of the task being performed.»

Infine il formato conversazionale è strettamente dipendente dall'architettura e dal tipo di iniziativa prevista per il sistema, la cui gestione può differire anche seconda dello stato del dialogo. Se il sistema è a iniziativa singola, questa può essere del sistema o dell'utente: nel primo caso solo il sistema pone domande all'utente al fine di completare il compito, nel secondo caso, l'iniziativa del solo parlante (generalmente usata nei sistemi per l'estrazione d'informazione e consultazione di un database) assegna al sistema il ruolo opposto nella conversazione. Nel caso dei sistemi a iniziativa mista il formato può configurarsi diversamente, essendo possibile per il parlante non dover seguire unicamente le sole indicazioni e domande del sistema. La struttura della conversazione può infine essere condizionato dal tipo di domande che il sistema pone all'utente: queste possono essere aperte, chiuse, o contenere indicazioni su quello che il parlante può dire o meno.

3.3 La semplificazione nel dialogo uomo-macchina

3.3.1 Studi empirici

Come affermato da Allen a proposito del dialogo pratico, un sistema di

dialogo può non possedere la stessa competenza conversazionale di un essere umano, per alcune ragioni strutturali: innanzitutto, il tipo di dialogo e l'attività da svolgere con l'interazione sfrondano la conversazione di molte delle funzioni pragmatiche e sociali normalmente sfruttate in una conversazione, oltre a incanalare lo svolgimento su un corso già definito; inoltre, il sistema, più che un interlocutore, è un'interfaccia funzionale a svolgere una determinata attività, l'utente dà perciò maggiore rilevanza a ciò che è sensibile per il sistema e per il compito, a scapito di altre funzioni non rilevanti; infine, nella semplificazione, svolgono un importante ruolo anche i preconcetti che un utente può avere sulle capacità conversazionali della macchina.¹⁹

La *vulgata* descrive il linguaggio usato nel dialogo uomo-macchina sostanzialmente come semplificato, come rivelano anche le varie denominazioni usate per etichettarlo: “formale” (Grosz 1977), “computer talk” (sul modello del *baby talk*, Guindon et al. 1986)²⁰, “telegrafico” (Guindon et al. 1987, Danieli 2004) e “computerese” (Reilly 1987). La natura semplificata del dialogo uomo-macchina è stata rilevata in numerosi studi su *corpora* (principalmente in inglese, ma anche in tedesco, svedese e francese), confrontati con gli equivalenti *corpora* uomo-uomo, facendo interagire due gruppi di parlanti, uno con un interlocutore uomo e uno con un sistema simulato (ruoli svolti dal medesimo sperimentatore con le medesime indicazioni, ad esempio Amalberti et al. 1993) oppure confrontando qualitativamente *corpora* ottenuti diversamente (ad esempio Fischer 2006).

Le denominazioni usate per il dialogo uomo-macchina tendono a evidenziare la scarsa varietà di strutture sintattiche, una morfosintassi scarna, un lessico omogeneo, l'assenza di segnali discorsivi, ecc.; un'interazione, quindi, gestita principalmente con comandi diretta e appoggiata principalmente alle indicazioni del sistema, più che un dialogo vero e proprio. Dal punto di vista sintattico è stata notata una sostanziale omogeneità e semplicità delle strutture, spesso riprese da quelle usate dal sistema (Hitzenberger – Whomser-Hacker 1995, Gustafson 2002), con una scarsa presenza di costruzioni interfrasali complesse. Ai pronomi, poco frequenti e il cui antecedente coreferente non è mai più distante di un turno (Guindon 1988, Dahlbäck – Jönsson 1989, Kennedy et al. 1988) vengono preferite le costruzioni ellittiche (Dahlbäck 1992); sono inoltre poco frequenti congiunzioni e preposizioni (Krause 1992). Gli unici due studi sull'italiano evidenziano da una parte la frequente

¹⁹ A questo argomento è dedicata la terza parte del capitolo (cfr. 3.4.2).

²⁰ *Computer talk* è anche usato in maniera più neutra, inteso come “all speech (and typed natural language utterances) directed towards an automatic speech processing system” (Fischer 2006).

omissione di elementi morfosintattici negli enunciati (pur mantenendo gli stessi contorni intonativi), come si può vedere negli esempi tratti da Danieli 2004:

- (1) Milano Torino domani ore dieci
- (2) Partenza Torino. Arrivo Bergamo. Domani dopo le dieci

Dall'altra la riduzione delle funzioni di alcuni meccanismi pragmatici come la ripetizione (Bazzanella et al. 1999). Nel lessico (a parte la naturale scrematura dovuta al dominio di applicazione del sistema di dialogo) viene registrata da più parti poca propensione alla variazione (Richards – Underwood 1984a, Hitzenberger – Whomser-Hacker 1995, Gustafson 2002). Da un punto di vista pragmatico e testuale, la struttura del discorso risulta semplificata per Amalberti et al. 1993, che rilevano un minor numero di connettivi e marcatori del discorso; così è anche per Krause – Hitzenberger 1992 e Hitzenberger – Whomser-Hacker 1995, che oltre a questa riduzione rilevano un minor numero di segnali discorsivi (segnali di feedback e di gestione del turno). Sempre sul piano pragmatico Morel 1989 e Johnston et al. 1994 riportano lo scarso uso di formule di cortesia e domande indirette, da legare anche alla scarsa attenzione (come vedremo in seguito) per funzioni conversazionali come il *grounding* (ad esempio, segnali discorsivi di feedback). Infine, da un punto di vista puramente quantitativo, Johnstone et al. 1994 e Oviatt 1995 rilevano un minor numero di parole per dialogo, mentre Amalberti et al. 1993 notano, al contrario, un maggior numero di parole per dialogo, che tende a decrescere con l'acquisizione di esperienza da parte del parlante.

Gli studi sul parlato uomo-macchina italiano viene riscontrata la stessa tendenza verso la semplificazione linguistica. Danieli 2004 parla di “parlato telegrafico”, facendo riferimento alle frequenti riduzioni su tutti i livelli linguistici: omissione di morfologia libera (articoli, preposizioni, pronomi); assenza di verbi; flessione verbale limitata; nominalizzazioni; giustapposizione di più enunciati: tutti fenomeni a cui corrisponde, tuttavia, il mantenimento dei contorni intonativi di delimitazione dell'enunciato. Il lavoro di Bazzanella et al. 1999 si incentra invece sul meccanismo della ripetizione: nel dialogo uomo-macchina, rispetto alle numerose funzioni che svolge nelle conversazioni quotidiane, la ripetizione ha principalmente il ruolo di correggere un'incomprensione da parte del sistema.

Oltre alle costrizioni imposte dal sistema e dal genere di attività svolto, che

determinano la semplificazione, anche i preconcetti degli utenti riguardo all'interazione con un sistema artificiale svolgono un ruolo importante. Amalberti et al. 1993 conducendo uno studio su due gruppi di parlanti che interagivano con il medesimo sistema simulato (un gruppo sapeva che il sistema era gestito da un essere umano, l'altro pensava di interagire con un sistema reale) non notano grandi differenze negli usi sintattici e lessicali, al netto delle differenze contestuali. Nell'analisi delle riformulazioni successive a un errore del sistema, Gieselmann – Waibel 2005 evidenziano la maggiore varietà di costruzioni sintattiche (sia a livello intrafrasale sia interfrasale) e di lessico. A livello pragmatico, sempre Amalberti et al. 1993 mettono in rilievo, a livello pragmatico, la simile presenza di richieste di conferma nei due gruppi, che denota la stessa preoccupazione per la reciproca comprensione con il sistema. Precedentemente, Richards – Underwood 1984b avevano notato l'uso di formule di cortesia quando queste venivano usate anche dal sistema; così Ward – Heeman 2000 avevano rilevato l'uso dei feedback di accordo nell'interazione con un sistema che li sfruttava a sua volta per segnalare la comprensione.

3.3.2 I modelli proposti per la semplificazione

Nella maggior parte degli studi empirici la tendenza alla semplificazione appare evidente (e prevedibile); d'altra parte, soprattutto dal punto di vista pragmatico (attenzione per il *grounding*, formule di cortesia, domande indirette, ecc.) e della strutturazione del discorso, la produzione linguistica del parlante è molto influenzata da quella del sistema, oltre ad essere “costretta” dall'architettura, dall'attività e dal dominio. Partendo da queste tendenze sono state proposte alcune classificazioni.

3.3.2.1 Un sottocodice o una varietà con differenze funzionali

Non avendo alcuni tratti dello standard, possedendo peculiarità proprie ed essendo ricorrenti alcune caratteristiche, Kraus – Hitzenberger 1992 hanno avanzato la proposta che il dialogo uomo-macchina sia un sottocodice. A sostegno di questa tesi stilano una lista di tratti che lo caratterizzerebbero come tale: semplificazione delle strutture sintattiche; minor numero di connettivi testuali, di pronomi, di formule

di cortesia, di allocutivi; tendenza a non esplicitare la struttura del discorso (e quindi del compito da svolgere) e diminuzione dei segnali discorsivi in generale (come feedback o segnali di presa del turno). In quanto sottocodice, le suddette caratteristiche possono essere prevedibili quando si presenta la situazione comunicativa uomo-macchina: indipendentemente dalle caratteristiche del sistema, dalle caratteristiche del parlante e dal compito da svolgere.

Per Johnstone et al. 1994, oltre alla natura del dialogo pratico, la differenza principale tra i due tipi di dialogo (uomo-uomo e uomo-macchina) è la non rilevanza, nel secondo, di funzioni della conversazione quotidiana come la cortesia e l'attenzione per il grounding. La mancanza di queste è da ricondurre alla scarsa rilevanza funzionale per la macchina di tali funzioni, secondo i presupposti del parlante umano. La diminuzione dei feedback, dei segnali di presa del turno, di formule di cortesia e di domande indirette (alcune di queste, come visto, rilevate anche da Krause – Hitzenberger 1992) sono quindi una conseguenza di tale presupposto. A sostegno di questa tesi gli autori hanno condotto un esperimento di tipo quantitativo su due *corpora* confrontabili: il risultato è una sostanziale equivalenza del numero di parole per turno tra i due tipi di dialogo, una volta esclusi dal corpus di dialogo uomo-uomo i segnali di feedback e le formule di cortesia. Per Fischer 2006, che pur critica questo modello, la scarsa preoccupazione per la cortesia può essere confermata invece, più che dalle formule esplicite, presenti anche nel dialogo uomo-macchina, dalla maggiore presenza delle forme imperative rispetto ai dialoghi uomo-uomo.

3.3.2.2 La rappresentazione dell'utente e la rappresentazione dell'interfaccia

Già Amalberti et al. 1993 assegnavano un ruolo importante alla rappresentazione del sistema da parte dell'utente: rappresentazione che per Fischer 2006 non è prevedibile ed è in continua negoziazione al variare dei parlanti. Nel suo studio, evidenziando la varietà delle interazioni a seconda degli utenti, sostiene l'assenza di proprietà specifiche, omogenee e prevedibili nel dialogo uomo-macchina, rilevandone invece la stretta connessione con le strategie che vengono di volta in volta messe in atto in base all'idea che si ha della macchina. Per Fischer, pur rimanendo funzionali per determinati contesti e pur portando, statisticamente,

risultati che possono essere utili nella progettazione, gli altri modelli hanno dei punti deboli. Per via della forte variazione tra parlanti, il dialogo uomo-macchina non può essere considerato un sottocodice: i parlanti non si comportano allo stesso modo²¹. Non si configura neanche come dialogo formale, privo di alcune caratteristiche della conversazione: l'attenzione per la comprensione del sistema è dimostrata dalle frequenti richieste di conferma e dalla varietà delle riformulazioni, su diversi livelli linguistici, nel caso in cui la macchina mostra di non aver compreso (cfr. 6.4); anche la cortesia è presente, ed è più frequente quando anche il sistema ne fa uso.

Per Fischer il dialogo uomo-macchina si configura come una negoziazione, che parte dai pregiudizi del parlante umano prima dell'interazione, pregiudizi che variano con l'acquisizione di conoscenze sulla macchina, sul linguaggio e sul compito. La macchina, inizialmente, può non essere ritenuta in grado di comprendere il linguaggio naturale, poiché l'interlocutore la immagina programmata per ragionamenti logici e quindi più capace di comprendere un linguaggio formale (cfr. 4.3). Successivamente i giudizi e il comportamento variano in seguito al confronto sul linguaggio: il parlante varia di volta in volta le strategie di riformulazione, cambiando la propria produzione su tutti i livelli (dalla prosodia alla sintassi). Anche il compito da svolgere influisce sul comportamento: il parlante organizza o esplicita con marcatori l'obiettivo e i sotto-compiti a seconda di quanto la macchina si mostra esperta dell'attività da svolgere (una auto-presentazione della macchina con una domanda aperta, ad esempio, porta il parlante a presupporre che la macchina conosca bene il contesto) e si mostra in grado di saper raggiungere un obiettivo tramite sotto-obiettivi. La negoziazione è inoltre evidente nella tendenza del parlante ad adattare la propria lingua a quella della macchina (come vedremo a proposito delle situazioni d'errore): in base a quello che la macchina dice e in base a quello che la macchina comprende (il parlante tende infatti a ripetere forme che la macchina mostra di aver compreso). Fischer associa quindi il dialogo uomo-macchina alla comunicazione interculturale: non omogeneo e strutturato, ma caratterizzato da strategie di negoziazione che variano di volta in volta a seconda dei pregiudizi del parlante e di come la macchina è in grado di cambiarli.

²¹ I segnali discorsivi di feedback, ad esempio, in base ai risultati di Fischer 1997, sembrano diminuire per quanto riguarda i parlanti maschili, mentre la stessa tendenza non si riscontra nelle parlanti femminili.

3.4. Il linguaggio naturale

L'interpretazione della situazione da parte degli utenti, che coinvolge l'interlocutore automatico, l'attività e il compito da svolgere, le esperienze avute con altre interfacce, porta a rivedere gli obiettivi della *human-likeness* (somiglianza a un interlocutore umano) e del linguaggio naturale²², normalmente visti come obiettivi per migliorare l'usabilità di un sistema.

Originariamente, la capacità di gestire il linguaggio naturale è nata come alternativa al linguaggio di programmazione necessario per interagire con un sistema: parallelamente si sono diffusi altri tipi d'interfaccia (principalmente grafiche) e meno dipendenti da un linguaggio (naturale o di programmazione). Ad oggi, come notano Edlund et al. 2007, naturalezza e *human-likeness*, applicate al linguaggio, hanno come valenza generica quella di essere “somiglianti alle capacità di comprensione e produzione proprie dell'essere umano”, valenza che, non tenendo conto della situazione comunicativa (del proprio interlocutore e dell'attività) può essere applicata solo per alcuni utenti e per alcune attività da svolgere, non incontrando invece il modo di interagire di altri parlanti e non presentandosi come soluzione migliore per svolgere alcune attività. Da un punto di vista applicativo la capacità di gestire il linguaggio naturale deve essere collocata nell'ottica delle interazioni linguistiche uomo-macchina, che con le interazioni uomo-uomo condividono solo alcune caratteristiche.

Se da una parte è vero che il NLP punta alla maggiore usabilità dei sistemi, dall'altra, un sistema che rassomigli il più possibile a un interlocutore non assicura automaticamente a una maggiore usabilità per ogni utente. La comprensione deve essere programmata per consentire l'interazione al posto di ogni possibile modo di interagire degli utenti, prevedendo contestualmente quello che può succedere in ogni punto del dialogo, da un punto di vista linguistico e pragmatico.

3.4.1 Human-like vs Tool-like

Le interfacce vocali sono polarizzate tra due modelli di interfaccia che ripropongono, in parte, la stessa classificazione che suddivide le interfacce a

²² Così come fa Allen a proposito del tipo di dialogo generalmente intrattenuto con un sistema artificiale, che comporta una riduzione di funzioni pragmatiche e conversazionali.

manipolazione diretta (*direct manipulation interfaces*) e le interfacce agenti (*agent interfaces*), altrove chiamate *tool-like interfaces* e *antropomorphic interfaces* (o *human-like interfaces*). Come precisa Qvartford 2004, la distinzione tra i due tipi di interfaccia è in realtà molto sfumata: ad esempio, la differenza tra *tool-like* e *human-like* non può essere unicamente ricondotta a un'ipotetica differenza tra *azione* e *linguaggio*; come insegna la pragmatica, con il linguaggio si possono compiere azioni, e non c'è grande differenza, tra dire “apri la porta” a un sistema domotico e spingere un bottone la cui funzione è aprire una porta: le interfacce vocali, generalmente “più umane” di quelle grafiche, possono avere caratteristiche più o meno umane, e tendere verso il modello *tool-like*. D'altra parte, anche con le interfacce *tool-like* l'interazione può coinvolgere l'aspetto linguistico: scegliendo da un menu a scorrimento una delle opzioni (presenti come testo) o salvando e nominando un file.

Ad oggi il polo delle interfacce *human-like* gode di un prestigio maggiore; tuttavia, per molti sistemi, soprattutto commerciali, è ancora preferito l'approccio *tool-like*, senz'altro più robusto e conosciuto, anche se più costrittivo. Le due tipologie sono dunque da intendersi come set di caratteristiche che possono appartenere in misura maggiore o minore a una determinata interfaccia.

3.4.1.1 Tool-like

Le interfacce *tool-like*, come per esempio quelle grafiche, si configurano come rappresentazioni di un oggetto su cui l'utente compie azioni fisiche, azioni che hanno un effetto visibile e immediato e che possono essere annullate. La forza delle interfacce *tool-like* è proprio il controllo degli oggetti (come le icone sul desktop), che consente di operare su una rappresentazione piuttosto che direttamente sul linguaggio di un programma. All'interno dello spazio definito dagli sviluppatori, l'iniziativa è completamente nelle mani dell'utente; pertanto l'apprendimento è intuitivo, difficilmente si presentano errori dovuti a un'errata interpretazione del comando da parte del sistema, il feedback è immediato ed è possibile annullare un'operazione. Al contrario, maggiore è il controllo da parte del sistema, maggiore è la difficoltà, per l'utente, di interpretare lo stato di un'operazione, per via della diminuzione di feedback: le operazioni “nascoste” e non esplicitate rendono maggiormente difficile l'apprendimento delle funzioni e della gestione, aumentando

la possibilità di errori, e rendono più complessa la reversibilità di operazioni su cui non si ha il continuo controllo.

3.4.1.2 Human-like

Le interfacce human-like più che rappresentare e simulare spazi fisici e oggetti su cui eseguire le azioni, rappresentano e simulano l'interazione che avviene tra persone. Basandosi sulle teorie pragmatiche dell'interazione tra persone, le interfacce human-like puntano sulla cooperazione tra i due agenti più che sul controllo totale del soggetto all'interno di uno spazio definito dal sistema. In genere, i sistemi di dialogo sono concepiti come interfacce human-like (per quanto, come vedremo, con diverse gradazioni, di cui alcune più vicine alle interfacce tool-like).

Il vantaggio di tali interfacce è legato strettamente all'uso del dialogo umano come spazio di interazione: l'utente, così, non ha bisogno di apprendere alcun linguaggio e possiede maggiore iniziativa, meno costretta dai parametri definiti dagli sviluppatori (ad esempio, in un'ipotetica interfaccia vocale per il computer è senz'altro più veloce dare il comando “apri il documento Tesi”, piuttosto che cercare il documento attraverso cartelle o caselle di ricerca).

Poiché sono sistemi concepiti come agenti in grado di cooperare, uno dei caratteri fondamentali delle interfacce human-like è la continua presenza di feedback di ciò che è stato compreso o non compreso, come avviene nelle conversazioni tra persone. Per cooperare, gli interlocutori devono agire su un terreno comune (*common ground*, cfr. Stalnaker 1978) in parte presupposto sulla base di assunzioni condivise, in parte continuamente ridefinito e ampliato nel corso della conversazione (*accumulation* o *grounding*): in altri termini, in una conversazione il partecipante oltre a capire ciò che l'altro partecipante ha detto, deve mostrare di aver capito, che l'enunciato è entrato a far parte del terreno comune e che la conversazione può continuare (Clark – Schaefer 1989). Il presupposto è che ogni azione compiuta (verbale o no) necessiti la comunicazione del suo effetto: sulla base di tale presupposto Clark (1996) ha ripreso il Principio di Chiusura elaborato da Norman (1988)²³: gli agenti che compiono un'azione necessitano dell'evidenza del successo dell'azione. Allo stesso modo, gli interlocutori segnalano se qualcosa è andato storto

²³ La presenza di un feedback è una delle regole fondamentali del design: quando non si può vedere direttamente l'effetto di una nostra azione con uno strumento, è necessario che qualcosa (un suono, una spia, ecc.) ci segnali che la nostra azione ha avuto successo, cfr. Norman 1988.

nell'azione (come ad esempio la comprensione reciproca) e se questa non ha avuto successo. L'acquisizione dell'enunciato all'interno della conversazione avviene tramite due fasi: presentazione e accettazione. La presentazione coincide con l'atto linguistico compiuto dal parlante; con l'accettazione il parlante successivo segnala che il contributo del parlante precedente è stato acquisito. Generalmente vengono usate cinque strategie per mostrare che l'enunciato è stato acquisito e che fa parte del terreno comune (Clark – Schaefer 1989): attenzione continuata (l'ascoltatore continua a seguire il turno senza interrompere), contributo successivo rilevante (l'ascoltatore/parlante successivo dà un contributo pertinente e rilevante rispetto al contributo del parlante precedente), accettazione esplicita (l'ascoltatore usa segnali di feedback durante il turno dell'interlocutore), dimostrazione (l'ascoltatore/parlante successivo riformula parzialmente il turno precedente, o lo completa), esposizione (l'ascoltatore/parlante successivo ripete interamente o in parte il contributo del parlante precedente). Tale classificazione di gestione delle acquisizioni, come vedremo a proposito delle situazioni d'errore, è ampiamente usata come modello nel campo delle interfacce vocali.

Tuttavia, per quanto possano esserci continui feedback, l'interazione verbale con un'entità non conosciuta porta l'utente a interrogarsi su come il sistema operi e processi i comandi, traendo conclusioni che spesso non coincidono con quelle effettive: questo può portare a effetti di sopravvalutazione (il sistema può capire qualsiasi cosa) o, paradossalmente, al non saper cosa dire. Per questo un'interfaccia dovrebbe essere *abitabile* (concetto elaborato in origine da Watt 1968), ossia dovrebbe manifestare all'interlocutore, oltre a ciò che ha capito, anche le sue capacità di comprensione.

3.4.2 La rappresentazione dell'utente

Le caratteristiche intrinseche delle varie interfacce non determinano da sole l'interazione: ciò che avviene durante il turno del parlante umano e come si configura la conversazione è anche legato alla rappresentazione che l'utente costruisce dell'interfaccia, che può non coincidere con la rappresentazione voluta dagli sviluppatori.

Ad esempio, per Nass l'uomo tende ad antropomorfizzare la macchina, quando questa presenta caratteristiche umane (Reeves – Nass 1996; Nass – Moon

2000; Nass – Brave 2005), comportandosi nello stesso modo in cui si comporterebbe con un uomo (Nass parla di “noncuranza”). Alcuni esperimenti sul comportamento (riportati in Nass – Moon 2000) danno come risultato l'applicazione delle medesime categorie sociali umane, come il genere sessuale e l'etnia; la presenza di meccanismi interazionali come la cortesia e la reciprocità (cioè la tendenza ad essere più disponibili verso chi ci ha precedentemente aiutato); la preferenza per sistemi artificiali con una “personalità” simile (rimando sempre a Nass – Moon 2000 per l'illustrazione degli esperimenti). Al contrario, le componenti tool-like, come ad esempio una maggiore iniziativa del sistema e una minore libertà al parlante, possono portare l'utente rappresentare l'interfaccia vocale allo stesso modo di un'interfaccia grafica.

3.4.3 Interfaccia o interlocutore?

Se è vero che la presenza di alcune caratteristiche (tool-like e human-like) può stimolare in misura maggiore o minore la vicinanza a una delle due tipologie, molto dipende da come il parlante interpreta il proprio interlocutore automatico e la situazione comunicativa in cui si trova.

Gli utenti di un sistema di dialogo “percepiscono” il sistema metaforicamente: in questo senso, le metafore ci aiutano a interpretare un oggetto non conosciuto, come un computer. Queste possono essere deliberatamente volute dai designer (come ad esempio l'ambiente desktop) oppure costituire un'interpretazione dell'utente, frutto della analisi della situazione comunicativa in cui si trova e di quelle in cui si è trovato precedentemente (sia con altri parlanti sia con altre interfacce)

Nel campo dei sistemi di dialogo la metafora è un interlocutore umano con cui cooperare come si fa quotidianamente con altri interlocutori umani; tuttavia l'interpretazione, basata su componenti più o meno umane, può spingere l'utente ad appoggiarsi ad altre metafore già conosciute, come ad esempio quella dell'interfaccia grafica. In alcuni studi è stata notata una differenziazione tra due stili di interazione: Riccardi – Gorin 2000 notano una polarizzazione della durata delle risposte ai saluti del sistema; Fischer 2006a e 2006b parla di utenti attivi e utenti non-attivi; Gustafson – Bell 2000 notano una differenza tra utenti che rispondono ai saluti del sistema e utenti che riducono al minimo le componenti pragmatiche della conversazione uomo-

uomo; anche Gonzalez-Ledesma – Moreno-Sandoval 2005) individuano due strategie opposte di interazione con il sistema: machine-like e human-like. Sul medesimo corpus C-ORAL-ROM (italiano) è stata condotta un'analisi (cfr. capitolo 4) in cui si evidenzia come le diverse strategie possano essere collegate anche alla gestione dell'iniziativa da parte del sistema e dell'utente. Per questo, Edlund et al. 2007 hanno avanzato l'ipotesi che gli utenti abbiano a disposizione due metafore per interagire con il sistema: quella della interfaccia e quella dell'interlocutore. Tramite la metafora dell'interfaccia grafica, gli input vocali sono equivalenti a input dati con il mouse: scegliere, ad esempio, il promemoria per un medicinale consiste nel selezionare la data e un orario da un formulario e scrivere il nome del medicinale. Al contrario, con la metafora dell'interlocutore, non possiamo parlare di input nel senso precedente, ma piuttosto di uso del linguaggio con un interlocutore con capacità conversazionali e linguistiche simili a quelle dell'essere umano, anche se solo parzialmente.

La doppia interpretazione non va naturalmente intesa, in fase di analisi, come una tipologia rigida a cui corrispondono alcuni fenomeni linguistici e conversazionali che ne escludono altri: rimane infatti d'altra parte il sistema, il cui comportamento linguistico può facilitare l'una o l'altra interpretazione e al contempo favorire la presenza di alcuni dei fenomeni.

Generalmente, i sistemi di dialogo concepiti come interfacce vocali sono sviluppati sulla base di interfacce già esistenti (come un centralino a toni o un'applicazione sul telefonino con alcuni dati da inserire vocalmente), presentando la medesima architettura e cambiando unicamente nella modalità di input dei dati da inserire.²⁴ Rispetto alle equivalenti interfacce hanno come vantaggio la possibilità di essere usati con mani e occhi occupati (ad esempio un sistema di dettatura); possono inoltre essere usati al posto di un'altra interfaccia del sistema occupata o nel caso di particolari disabilità.

Un sistema di dialogo concepito come interlocutore ingloba le caratteristiche di un'interfaccia vocale, con ulteriori caratteristiche che potremmo definire umane: consente una gestione “intelligente” degli input del parlante, il quale è più libero di scegliere come presentarli; supporta varie strategie di gestione degli errori in relazione alla certezza che ha dell'acquisizione; sfrutta alcune funzioni linguistiche

²⁴ Ad esempio, spesso il turno è definito manualmente (tramite un pulsante *push-to-talk*) o viene rilasciato dal sistema (con un segnale sonoro).

(memoria del contesto, strategie di feedback, ecc.) e conversazionali, come ad esempio la presa e la cessione del turno (che, al contrario, è tutt'altro che umana quando gestita con il meccanismo del push-to-talk).

I sistemi concepiti come interfacce sono molto più diffusi dei secondi, soprattutto a livello commerciale; i secondi, al contrario, sono sviluppati principalmente sviluppati in ambito universitario e di ricerca, anche per motivazioni strutturali: se le interfacce vocali possono di fatto sostituire altre interfacce (ad esempio, quelle a toni dei centralini telefonici), un sistema di dialogo concepito come interlocutore non può ancora sostituire completamente un uomo nello svolgimento di un'attività, ma al massimo riprodurre alcune funzioni, rassomigliando sempre più a un interlocutore umano.

Come notano Edlund et al. 2006, non è detto che la scelta di una delle due metafore coincida con quella dell'utente: l'interazione può essere maggiormente problematica se l'utente ha aspettative sulle capacità del sistema differenti rispetto a quelle che il sistema è in grado di gestire (cfr. Tab 2).

		intended metaphor (system)	
		human	inteface
expected metaphor (user)	human	D ₄ S ₁ Hi there!	D ₅ S ₁ This is X-system. Where do you want to go?
		D ₄ U ₁ Hi.	D ₅ U ₁ Hi. I'm not sure – is England nice this time of year?
		D ₄ S ₂ Where would you like to go?	D ₅ S ₂ Please state where you want to go.
		D ₄ U ₂ To London, please.	
	inteface	D ₆ S ₁ Hi there!	D ₇ S ₁ This is X-system. Where do you want to go?
		D ₆ U ₁ <silence>	D ₇ U ₁ London.
		D ₆ S ₂ Hello?	D ₇ S ₂ To London.
		D ₆ U ₂ Oh - hi.	<pause>
		D ₆ S ₂ Where would you like to go?	From where?
		D ₆ U ₂ To London, please.	

Tab. 2: Possibili interazioni in relazione alla metafora usata dall'utente e dal sistema (da Edlund et al. 2006)

La differenza di rappresentazioni tra parlante e sistema può essere causata dal sistema stesso, nel caso sia rappresentato da entrambe le metafore in punti diversi del dialogo (ad esempio, il sistema SIRI, che a una domanda aperta iniziale fa seguire operazioni gestibili solo tramite command-and-control), o simuli delle funzioni che in realtà non può gestire (ad esempio, l'uso di feedback o di uno stile conversazionale umano, fatto per rendere accattivante il sistema). Porzel 2006 ha evidenziato una scarsa attenzione per i moduli di generazione del linguaggio, spesso considerati come

la parte meno problematica di un sistema, ma la cui importanza è fondamentale per aiutare l'utente a rappresentare correttamente l'interfaccia.

3.5 Sommario e conclusioni

Il dialogo uomo-macchina possiede alcune proprietà peculiari tra i possibili generi di dialogo. A livello contestuale, la rilevanza di alcuni tratti (come spazio e tempo) è definita dalle caratteristiche del sistema, il quale, inoltre, governa parzialmente o totalmente lo svolgimento dell'interazione. Lo stesso dialogo tra parlante e sistema, generalmente, si caratterizza rispetto ad altri generi dialogici per essere fortemente focalizzato sullo svolgimento di un'attività. A un livello linguistico, la semplificazione e la riduzione dipendono innanzitutto dai limiti imposti dal sistema; lo stesso avviene a livello pragmatico e conversazionale. Tuttavia, lo spazio concesso dal sistema non sempre viene sfruttato completamente, mentre altre volte viene forzato al di là delle capacità del sistema; l'utente, infatti, tende a interagire con il sistema rifacendosi a rappresentazioni già conosciute: l'interfaccia grafica e l'interlocutore umano. Oltre alle precedenti esperienze che un utente può aver avuto con altre interfacce, anche la rappresentazione che il sistema dà di sé può condizionare il comportamento dell'utente: un sistema più “umano” favorisce una minore preoccupazione dell'utente per la forma con cui presentare i propri input, mentre un sistema marcatamente “meccanico” può stimolare un comportamento più cauto e sintetico. È quindi importante che la rappresentazione del sistema sia coerente con le sue capacità di comprensione, con l'attività che deve svolgere e con il target di utenti a cui si rivolge.

4. Analisi del *corpus* C-ORAL-ROM

4.1 Introduzione

Nel presente capitolo viene analizzato un *corpus* di interazioni uomo-macchina raccolto nell'ambito del progetto C-ORAL-ROM (Cresti et al. 2004). Il *corpus* analizzato è costituito da 46 conversazioni²⁵ tra soggetti sperimentali chiamati appositamente per la raccolta e un sistema automatico di informazioni sugli orari ferroviari, progettato per il corpus C-ORAL-ROM. Nel *corpus* spagnolo sono stati individuati due tipologie di interazione da parte degli utenti (Gonzalez-Ledesma – Moreno-Sandoval 2005): una definita *machine-like* (l'utente interagisce con il sistema come con un'interfaccia grafica) e un altro definito *human-like* (l'utente interagisce con il sistema come con un interlocutore). La stessa suddivisione può essere riscontrata nel *corpus* italiano: a differenza del lavoro di Gonzalez-Ledesma – Moreno-Sandoval 2005 si tenterà di ricondurre i due diversi comportamenti alla gestione dell'iniziativa; per cui, oltre alla descrizione e all'analisi linguistica del *corpus*, si proverà a correlare la *facies* linguistica all'iniziativa da parte del parlante nella conversazione, allo scopo di individuare potenziali profili di interazione, con possibili risvolti applicativi.

4.2 Situazione e tipo di interazione

4.2.1 Situazione extra-linguistica

L'interazione avviene tramite telefono e quindi unicamente per via vocale (come avviene d'altronde con la maggior parte dei sistemi di dialogo non multimodali). Le conversazioni avvengono in tempo reale, tuttavia non sono presenti sovrapposizioni e segnali di feedback durante il turno del sistema: bisogna evidenziare, infatti, che il sistema non consente il *barge-in* (possibilità di

²⁵ Il *corpus* originario è costituito da 51 conversazioni: sono stati esclusi cinque dialoghi in cui non avviene una vera e propria interazione (in un caso il parlante sbaglia numero, in tre casi indica solo la città di partenza senza proseguire l'interazione, in un altro caso, dopo l'output del sistema non c'è alcuna risposta).

interrompere il sistema sovrapponendosi, cfr. 6.3.3.3) e che non è registrato il comportamento del parlante durante il turno del sistema. Il sistema è progettato per consentire l'iniziativa mista; il parlante non è unicamente costretto da ciò che chiede il sistema, ma può rispondere liberamente e ha la possibilità di organizzare l'interazione e la risoluzione del compito. In relazione alla situazione va inoltre segnalata la presenza occasionale di altre persone in prossimità del parlante, rendendo frequenti fenomeni di *off-talk* (Batliner et al. 2006), ovvero di parlato non indirizzato alla macchina (esclusi naturalmente dall'analisi).²⁶

*MAC: sono stati selezionati undici treni da Prato a Milano Centrale / con partenza lunedì trenta dicembre // è stato selezionato un solo treno locale da Prato a Milano Centrale / con partenza lunedì trenta dicembre / verso le uno e undici // treno uno / treno locale numero duemilaseicentocinquantanove / con vagone ristorante / partenza alle due e sette / arrivo alle cinque e cinquantatre // posso fare altro per lei ?

*WOM: sì // **alle due la notte / e si arriva alle cinque la mattina** //

4.2.2 Struttura dell'interazione

Le interazioni rientrano tra i dialoghi pratici per via del forte orientamento sullo svolgimento di un'attività (avere delle informazioni sugli orari ferroviari) e per la netta delimitazione del dominio (cfr. 3.2.2; Allen et al. 2000; Leech – Weisser 2005).

Ogni interazione inizia con una breve introduzione in cui il sistema illustra le possibilità di interazione del parlante, che può formulare a piacimento le proprie richieste o limitarsi a rispondere alle domande del sistema. L'attività può essere divisa in due fasi:

1° fase: fornire i dati sul viaggio che si intende intraprendere

2° fase: una volta segnalata dal sistema la quantità di treni corrispondenti alla richieste, scegliere i treni su cui avere maggiori informazioni

²⁶ Le trascrizioni degli esempi sono state riprese dal *corpus* C-ORAL-ROM e ne seguono i criteri di trascrizione: il turno del sistema è segnalato con MAC, quello dei soggetti sperimentali da MAN e WOM, rispettivamente, se sono un uomo o una donna. Le pause brevi sono segnalate da /, mentre quelle lunghe da //. Le esitazioni da &hes.

4.2.2.1 Apertura: all'inizio di ogni interazione viene introdotto il sistema e vengono richieste la città di partenza e la data; inoltre, il sistema comunica all'utente le diverse strategie che possono essere usate per l'interazione.

*MAC: Buongiorno // servizio automatico di informazioni su orari ferroviari // può rispondere alle domande una per una / oppure esprimere una richiesta con una frase complessa // dica quando e da che città vuole partire //

4.2.2.2 Negoziazione: dopo ogni input il sistema chiede conferma dei dati acquisiti (che possono essere relativi all'attività o relativi all'interazione stessa, come l'intenzione di interrompere la conversazione o di cancellare tutti i dati).

*MAC: vuole partire da Prato il giorno lunedì trenta dicembre?

*MAC: vuole terminare l' interazione ?

Nel caso l'acquisizione non sia confermata o manchino alcune informazioni necessarie (città di partenza, città di arrivo, data), il sistema può richiedere i dati mancanti con richieste di completamento:

*MAC: dica quando e da che città vuole partire //

*MAC: dica la stazione di arrivo //

4.2.2.3 Chiusura/apertura: il sistema, una volta individuata la data e le città di partenza e destinazione²⁷, fornisce il numero di treni disponibili; nello stesso turno, apre la seconda fase chiedendo quali treni l'utente vuole selezionare²⁸:

*MAC: sono stati selezionati undici treni da Prato a Milano Centrale / con

²⁷ Oltre a questi dati necessari, il parlante può facoltativamente restringere la ricerca indicando il tipo di treno, la parte della giornata o l'ora, la presenza o meno di un vagone ristorante, ecc.

²⁸ Questo turno del sistema varia a seconda del numero di treni disponibili. Se i treni sono pochi, il sistema fornisce per ognuno le informazioni, se sono molti, il sistema indica solo il numero totale: in questi la richiesta successiva in cui si chiede di scegliere un treno non sembra molto appropriata, in quanto presuppone che il parlante sappia l'orario di ognuno dei treni.

partenza lunedì trenta dicembre // quali treni vuole selezionare?

4.2.1.4 Negoziazione: dopo la risposta dell'utente, il sistema chiede conferma dei treni richiesti:

*MAC: vuole usare un treno eurostar ?

Se non sono presenti i treni richiesti, chiede all'utente di modificare la richiesta²⁹:

*MAC: mi dispiace / non ho trovato nessun treno da Trento a Roma Termini con partenza sabato quattordici dicembre / tra le sette e le tredici // dovrebbe modificare la sua richiesta // vuole partire tra le sette e le tredici ?

4.2.2.5 Chiusura/(apertura): una volta individuato il treno (o i treni) il sistema presenta informazioni più dettagliate sul treno, chiedendo alla fine se si ha intenzione di proseguire la conversazione con nuovo compito.

*MAC: treno uno / treno locale numero duemilanovecentosettantaquattro / con solo servizio di prima classe / partenza alle sette e quarantanove / arrivo alle undici e ventisette // posso fare altro per lei ?

4.3 Successo

Come già notato in Gonzalez-Ledesma – Moreno-Sandoval 2005, è difficile identificare con precisione se alcuni dialoghi hanno avuto successo in un esperimento: il parlante, non essendo un utente realmente interessato al successo dell'interazione, a volte cambia intenzioni e obiettivo, per via delle difficoltà incontrate nell'interazione o per ragioni difficilmente ricavabili dalla conversazione. Tuttavia, a differenza dell'analisi sul *corpus* spagnolo, annovero tra i casi di successo

²⁹ Anche in questo caso l'output del sistema risulta talvolta equivoco: dopo aver segnalato la necessità di cambiare la richiesta, il sistema chiede comunque se si vuole continuare con la richiesta precedente, come se chiedesse conferma o desse un consiglio: *MAC: mi dispiace / non ho trovato nessun treno eurostar da Firenze Santa Maria Novella a Roma Termini con partenza lunedì nove dicembre // dovrebbe modificare la sua richiesta // vuol usare un treno eurostar”.

gli episodi in cui il parlante cambia obiettivo e l'attività viene svolta, non essendo possibile per la macchina eseguire qualcosa di diverso da ciò che le viene detto dal parlante.

Successo: 29 - 63,04%

Insuccesso: 17 – 36,96%

4.4 Iniziativa

Il sistema di dialogo è a iniziativa mista: per il parlante è possibile organizzare la conversazione prendendo l'iniziativa (il controllo che esercita un interlocutore sul turno successivo dell'altro interlocutore, cfr. 2.4.1) nel proprio turno o, al contrario, può lasciare l'organizzazione dell'interazione al sistema. A differenza dei sistemi a iniziativa del sistema, con cui il parlante può interagire unicamente rispondendo alle domande, nei sistemi a iniziativa mista è consentito correggere direttamente il sistema nel turno successivo a una richiesta di conferma, fornire dati rilevanti anche se non esplicitamente richiesti, cambiare lo stato del dialogo in qualsiasi punto della conversazione; il parlante ha, quindi, la possibilità di controllare l'interazione al pari del sistema (cfr. 2.4.1.2).

Come scritto nel capitolo 3, il comportamento dei parlanti che interagiscono con un sistema automatico di dialogo non è riducibile ad un'unica tipologia, ma oscilla su un *continuum* i cui estremi sono un approccio *human-like* (il parlante si comporta come se stesse interagendo con una persona) e uno *interface-like* (il parlante si comporta come se stesse interagendo con un' interfaccia grafica). L'iniziativa, definendo parzialmente chi conduce l'interazione, può essere un indicatore di collocamento su questo *continuum*. Nell'analisi, riprendendo Walker – Whittaker 1990, è stato considerato un passaggio di iniziativa, quando il parlante, nel suo turno, non si limita a rispondere alle variabili richieste dal sistema, ma ne presenta altre, che oltre a soddisfare la richiesta del sistema, incidono sulla sequenza dei turni.

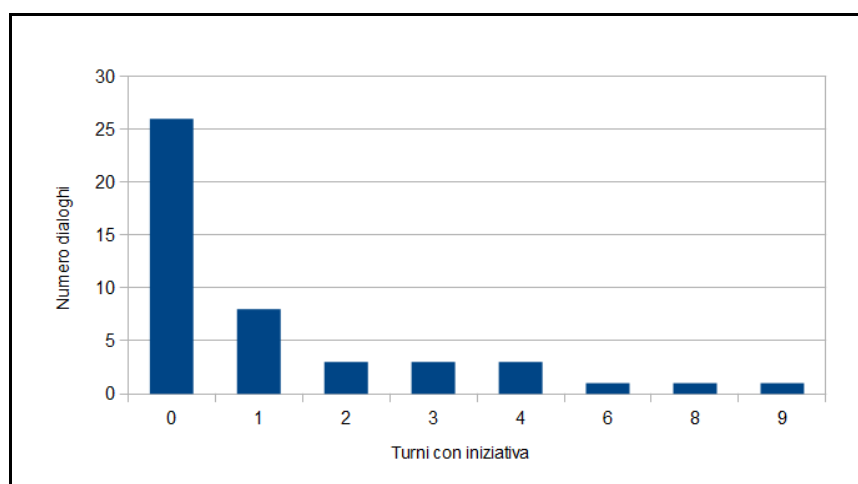
4.4.1 Gestione dell'iniziativa in C-ORAL-ROM

A fronte della possibilità per il parlante di prendere l'iniziativa della conversazione con il sistema, la percentuale di turni *user-initiative* è piuttosto bassa. Su 481 turni in cui è possibile prendere l'iniziativa, solo in 63 turni questo effettivamente avviene:

Iniziativa parlante: 13,10%

Iniziativa sistema: 86,90%

In più della metà dei dialoghi (26 dialoghi) non avviene alcun passaggio di iniziativa e il parlante si limita a seguire le richieste del sistema; nei restanti dialoghi il passaggio di iniziativa si presenta dal minimo di un turno (8 dialoghi) fino a 9 turni (in un dialogo). Nel grafico 1 sono presentati il numero di dialoghi (y) e il numero di turno con passaggio dell'iniziativa (x):



Graf. 1: Numero di dialoghi con turni di iniziativa del parlante

4.4.2 Posizione

Il fatto che il sistema renda possibile l'iniziativa mista, e che quindi l'interazione non sia rigidamente legata alla struttura impostata dal sistema, rende possibile il passaggio d'iniziativa successivamente a qualsiasi richiesta del sistema. Nel *corpus* analizzato la distribuzione non è omogenea: in più della metà dei casi l'iniziativa viene presa dopo una richiesta di conferma da parte del sistema, mentre in

minor numero dopo domande aperte (come la richiesta iniziale o quella finale), o richieste di completamento:

Passaggio di iniziativa dopo una...		
... richiesta di conferma	38	65,52%
... domanda aperta	12	20,68%
... richiesta di completamento	8	13,79%

Tab. 3: Distribuzione dei turni precedenti a un passaggio di iniziativa

4.4.3 Funzione

Il parlante può prendere l'iniziativa per intervenire su un turno precedente (*backward-looking*), per presentare più dati di quelli richiesti o formulare nuove richieste (*forward-looking*) o per chiudere la conversazione. Coerentemente con i dati precedenti, la maggior parte dei passaggi di iniziativa avvengono per correggere un errore precedente del sistema:

Funzione conversazionale		
Backward-looking	29	50%
Forward-looking	18	31,03%
Chiusura	11	18,97%

Tab. 4: Funzione dell'input quando avviene un passaggio di iniziativa

4.4.4 Tipologia delle funzioni

4.4.4.1 Backward-looking

Il turno del parlante dedicato a correggere il sistema dopo una richiesta di conferma (in cui il sistema mostra di aver acquisito informazioni sbagliate), è

generalmente costituito da un feedback negativo (in risposta alla diretta del sistema) e dall'informazione corretta:

*MAC: non ho capito // vuole arrivare a Ancona ?

*WOM: no // Roma //

Altrimenti, il parlante può riformulare nuovamente la richiesta completa, seguendo in parte la formulazione usata dal sistema (*alignment*, cfr. 4.6.2):

*MAC: vuole arrivare a Ancona ?

*WOM: no // voglio arrivare a Roma //

In altri casi viene ripetuta e rifiutata l'informazione errata, a cui segue la correzione:

*MAC: non ho capito . vuole partire da como-san-giovanni ?

*WOM: no / non da Como // da Trieste //

Altrimenti il parlante può limitarsi a presentare l'informazione corretta, senza segnalare esplicitamente l'errore del sistema:

*MAC: vuole partire da Enna ?

*WOM: Firenze //

4.4.4.2 Forward-looking

In un minor numero di casi il parlante prende l'iniziativa per introdurre dati non esplicitamente richiesti dal sistema o per presentare nuove richieste: il parlante può così parzialmente organizzare la risoluzione del compito:

*MAC: buongiorno // servizio automatico di informazioni su orari ferroviari // può rispondere alle domande una per una / oppure esprimere una richiesta con una frase complessa // dica quando e da che città vuole partire //

*WOM: voglio partire venerdì sera da Firenze / e andare a Genova //

L'iniziativa viene ceduta necessariamente quando, completato il compito, il

sistema chiede se si vuole formulare una nuova richiesta: il parlante può prendere o meno l'iniziativa a seconda dell'intenzione o meno di proseguire la conversazione, una volta finito il compito:

*MAC: è stato selezionato un solo treno da Milano Centrale a Napoli Centrale / con partenza martedì dieci dicembre / verso le ore dieci // treno uno / treno Eurocity numero tremilasettecentosessanta con servizio bar / partenza alle nove e tre / arrivo alle dodici e quindici // posso fare altro per lei ?

*WOM: vorrei / un treno / per il ritorno //

L'iniziativa può anche essere presa dopo una richiesta di conferma con lo scopo di proseguire con una nuova informazione, rispondendo in maniera affermativa e introducendo i nuovi dati:

*MAC: vuole partire da Roma Termini tra le sette e le tredici / il giorno mercoledì undici dicembre ?

*MAN: sì / va bene / e voglio andare a Firenze //

4.4.4.3 Chiusura

Il passaggio di iniziativa può avvenire inoltre quando il parlante decide di chiudere l'interazione:

*MAC: mi dispiace / non ho trovato nessun treno da Trento a Roma Termini con partenza sabato quattordici dicembre / tra le sette e le tredici // dovrebbe modificare la sua richiesta // vuole partire tra le sette e le tredici ?

*WOM: no grazie / arrivederci // basta così //

4.5 La gestione dell'iniziativa come risorsa per definire un profilo di interazione

Come illustrato nel capitolo 3, le conversazioni tra persone e sistemi artificiali differiscono dalle normali conversazioni tra persone: per diversa situazione

comunicativa, per le limitate capacità del sistema e per gli obiettivi e il dominio dell'interazione. Questo ha portato generalmente a considerare il dialogo uomo-macchina come una forma semplificata di interazione: scarsa varietà di strutture sintattiche, morfosintassi scarna, preferenza per le costruzioni ellittiche, scarsa presenza di segnali discorsivi e scarsa varietà di strategie pragmatiche. Tuttavia, per la forte variabilità a seconda dei parlanti, il dialogo uomo-macchina non si può considerare al pari di una varietà: la semplificazione dipende, infatti, in larga parte dall'attività del parlante, determinata dall'idea che questi ha della macchina, da come intende usarla e da come la macchina è capace di cambiare il suo comportamento (cfr. Edlund et al. 2008 per una breve rassegna). L'iniziativa può essere un indice per sapere quanto il parlante partecipa attivamente all'interazione? Esiste una correlazione tra iniziativa e *facies* linguistica, o tra iniziativa e strategie pragmatiche?

Per vedere eventuali correlazioni, il *corpus* è stato suddiviso in due sotto-*corpus*: un *corpus* con dialoghi in cui l'iniziativa rimane sempre nella mani del sistema e un *corpus* composto da dialoghi con almeno un turno di iniziativa da parte del parlante. La suddivisione è convenzionale: le dimensioni del *corpus* non consentono, infatti, lo studio dei differenti gradi di controllo dell'interazione; quindi, a fronte dell'omogeneità delle interazione nel sotto-*corpus* senza iniziativa, il sotto-*corpus* con almeno un turno di iniziativa potrebbe non rappresentare un unico tipo di interazione ma una proiezione di più strategie.

4.6 Analisi

4.6.1 Dialoghi senza iniziativa

Più della metà dei dialoghi (26 su 46) non presentano nessun passaggio di iniziativa: in questi casi la conversazione e la risoluzione del compito sono guidati unicamente dal sistema, come nel seguente dialogo riportato come esempio:

- *MAC: buongiorno // servizio automatico di informazioni su orari ferroviari // può rispondere alle domande una per una / oppure esprimere una richiesta con una frase complessa // dica quando e da che città vuole partire //
- *WOM: ventitre dicembre / da Catania //
- *MAC: vuole partire da Catania Centrale verso le ore venti ?

*WOM: no //

*MAC: dica quando e da che città vuole partire //

*WOM: ventitre dicembre / da Catania //

*MAC: vuole partire da Catania Centrale il giorno lunedì ventitre dicembre ?

*WOM: sì //

*MAC: dica la stazione di arrivo //

*WOM: Palermo //

*MAC: vuole arrivare a Palermo Centrale ?

*WOM: sì //

*MAC: sono stati selezionati tre treni da Catania Centrale a Palermo Centrale / con partenza lunedì ventitre dicembre // le leggo i dati principali // treno uno / partenza alle sette e quarantanove / arrivo alle otto e quaranta // treno due / partenza alle quindici e trentotto / arrivo alle sedici e sedici // treno tre / partenza alle ventitre e ventisette / arrivo alle zero e quarantuno di martedì ventiquattro dicembre // quali treni vuole selezionare ?

*WOM: il primo //

*MAC: treno tre / treno interregionale numero milleseicentosessantaquattro / con vagone ristorante / partenza alle ventitre e ventisette / arrivo alle zero e quarantuno di martedì ventiquattro dicembre // posso fare altro per lei ?

In questo gruppo di dialoghi possiamo identificare la tendenza, da parte del parlante, a non intervenire mai al di fuori delle richieste del sistema. In più della metà dei dialoghi (14) la morfosintassi è completamente assente (articoli, preposizioni, flessione verbale, ecc.): l'utente si limita a fornire i dati rilevanti per il sistema, eliminando qualsiasi elemento grammaticale o pragmatico "superfluo" o considerato non rilevante:

*MAC: buongiorno // servizio automatico di informazioni su orari ferroviari // può rispondere alle domande una per una / oppure esprimere una richiesta con una frase complessa // dica quando e da che città vuole partire //

*WOM: Firenze //

*MAC: vuole partire da Firenze Santa Maria Novella ?

*WOM: sì //

*MAC: dica che giorno vuole partire //

*WOM: venerdì //

*MAC: vuole partire il giorno venerdì tredici dicembre ?

*WOM: sì //

*MAC: dica la stazione di arrivo //

*WOM: Roma //

*MAC: vuole arrivare a Como San Giovanni ?

*WOM: no //

*MAC: dica la stazione di arrivo //

*WOM: Roma //

Nei restanti dodici si riscontra, complessivamente, la stessa impressione di Danieli 2004: una microsintassi molto scarna e una preferenza generale per le costruzioni ellittiche; viene quasi sempre evitata la formulazione completa della richiesta, presente solo in 5 dialoghi (complessivamente in 7 turni); in nessuno dei dialoghi, inoltre, sono presenti turni con costruzioni macrosintattiche che vadano oltre la singola proposizione.

La tendenza all'ellissi e alla coesione lasciata implicita si riscontra anche sul piano testuale: sono infatti generalmente assenti elementi anaforici che riprendono dati dell'interlocutore o del parlante stesso. Quando viene chiesto dal sistema quale treno scegliere, in genere il parlante si limita a ripetere il numero associato al treno (se la lista è già stata detta):

*MAC: sono stati selezionati tre treni da Prato a Milano Centrale / con partenza lunedì trenta dicembre // le leggo i dati principali // treno uno / partenza alle sette e quarantanove / arrivo alle undici e ventisette // treno due / partenza alle quindici e trentotto / arrivo alle diciannove e quarantasei // treno tre / partenza alle ventitre e ventisette / arrivo alle tre e venticinque di martedì trentuno dicembre // quali treni vuole selezionare ?

*WOM: uno //

Anche quando viene introdotto un elemento non indicato precedentemente dal sistema (ad esempio il tipo di treno), il turno viene unicamente occupato dall'elemento nuovo, senza nessuna ripresa testuale degli elementi già dati:

*MAC: sono stati selezionati dodici treni da Firenze Santa Maria Novella a Roma Termini con partenza sabato quattordici dicembre // quali treni vuole selezionare ?

*WOM: nel pomeriggio //

Sul piano pragmatico, la completa assenza di segnali discorsivi conferma il grado zero di gestione dell'interazione da parte del parlante. In caso di errore (risolto sempre secondo la sequenza richiesta di conferma – risposta negativa – nuova richiesta – ripetizione dell'informazione) il turno di riparazione è quasi sempre è la ripetizione³⁰ completa dell'enunciato:

Ripetizione totale: 43/47

*WOM: **trenta / dicembre / Montecatini //**

*MAC: vuole cancellare tutti i dati immessi finora ?

*WOM: sì //

*MAC: dica quando e da che città vuole partire //

*WOM: **trenta / dicembre / Montecatini //**

Riformulazione espansa: 1/47

*WOM: **vorrei partire da Genova //**

*MAC: grazie // vuole partire da Bari Centrale e arrivare a Piacenza ?

*WOM: no / **vorrei partire / da Genova / sabato mattina //**

Riformulazione ridotta: 3/47

*MAC: [...] dica quando e da che città vuole partire //

*MAN: **Firenze // venerdì //**

*MAC: vuole partire da Firenze Santa Maria Novella e arrivare a Biella San Paolo ?

*MAN: no //

*MAC: dica quando e da che città vuole partire //

³⁰ Per la tassonomia si riprende la classificazione di Bazzanella et al. 1999.

*MAN: **Firenze** //

La maggior parte dei dialoghi (69%), infine, oltre a non presentare naturalmente atti di chiusura della conversazione proposti dal parlante, si concludono con la chiusura del canale da parte del parlante, senza che venga segnalata l'intenzione di chiudere l'interazione (rispondendo negativamente alla domanda aperta conclusiva del sistema), come invece avviene nell'esempio:

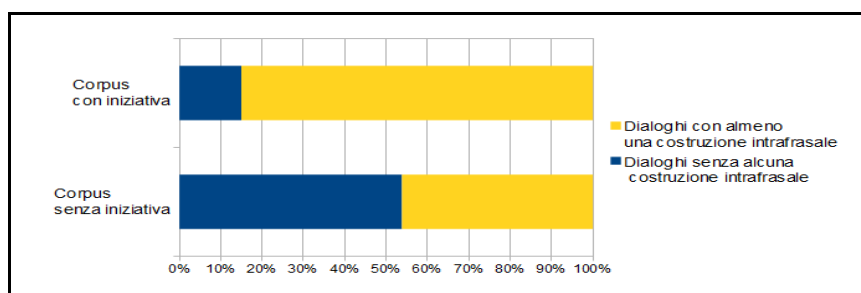
*MAC: treno uno / treno locale numero millenovecentodue con servizio bar / partenza alle quattordici e zero / arrivo alle sedici e venticinque // posso fare altro per lei ?

*WOM: no //

*MAC: grazie per aver usato questo servizio // a risentirci //

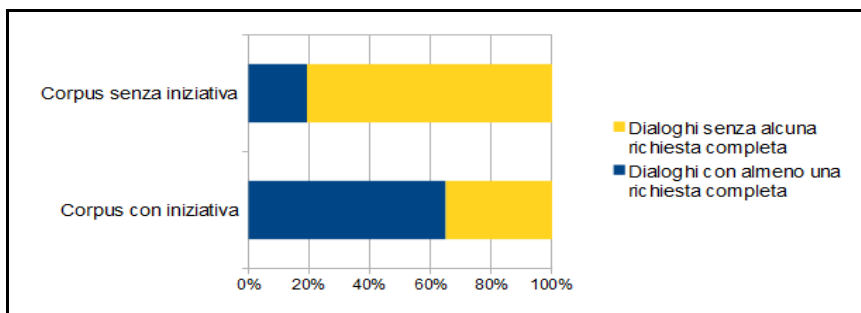
4.6.2 Dialoghi con iniziativa

Prima di confrontarlo con il sotto-*corpus* senza iniziativa appena visto, bisogna ricordare che il gruppo di dialoghi con almeno un turno di iniziativa da parte del parlante è costituzionalmente più eterogeneo, rappresentando un insieme di strategie diverse, che vanno da un minore a un maggiore grado di iniziativa. Non essendoci, nel *corpus* esaminato, un maggior numero di dialoghi con turni di iniziativa, possiamo comunque considerare l'insieme nella sua totalità, per verificare eventuali differenze linguistiche e pragmatiche, e tracciare almeno due profili di interazione. Tra i dialoghi con almeno un turno di iniziativa la propensione all'ellissi è minore; come indice si può vedere la rara presenza di dialoghi senza alcuna costruzione, a differenza del gruppo senza iniziativa, con cui sono comparati nel grafico 2:



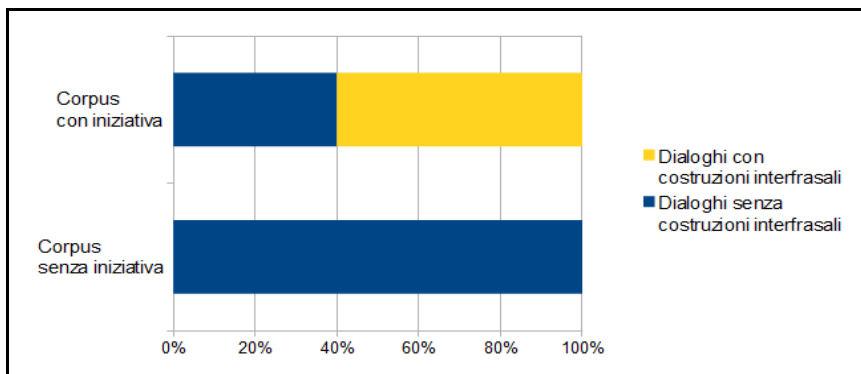
Graf. 2: Distribuzione di costruzioni intrafrasali nei due sotto-corpus

Il resto del sotto-*corpus* con iniziativa mostra una maggiore presenza e varietà di costruzioni intrafrasali rispetto al sotto-*corpus* senza iniziativa. Più della metà dei dialoghi (13 su 20) presenta almeno una formulazione completa di una richiesta, per un totale di 45 turni con richieste complete³¹:



Graf. 3: Distribuzione di richieste complete nei due sotto-corpus

È indicativa la presenza di costruzioni interfrasali (coordinate in 13 casi, subordinate in 3 casi) in quasi la metà dei dialoghi; anche in questo il sotto-corpus con iniziativa differisce da quello senza iniziativa, dove sono totalmente assenti.



Graf. 4: Distribuzione di costruzioni interfrasali nei due sotto-corpus

Sul piano testuale la coesione è più esplicita: per quanto dopo la richiesta di selezione del treno siano comunque presenti turni in cui viene selezionato il treno dalla lista indicandone unicamente il numero associato, nel *corpus* con iniziativa troviamo ripetizioni anaforiche intere:

*MAC: sono stati selezionati nove treni da Firenze Santa Maria Novella a Venezia Santa Lucia / con partenza mercoledì undici dicembre // quali treni vuole selezionare ?

*MAN: un treno per la tarda mattinata //+

O tramite un pronome:

*MAC: sono stati selezionati sette treni da Milano Centrale a Napoli Centrale / con partenza martedì dieci dicembre // quali treni vuole selezionare ?

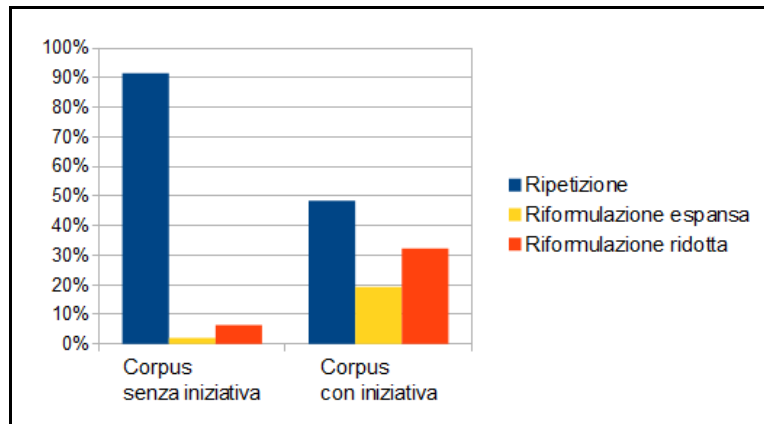
*WOM: &he / quelli verso le dieci della mattina //

Anche sul piano pragmatico incontriamo una maggiore varietà di forme e strategie. Nel sotto-*corpus* sono innanzitutto presenti segnali discorsivi (in 11 dialoghi), che svolgono funzioni interazionali base:

- Segnali di chiusura: 9 (*basta così, va bene così, a posto così*)
- Segnali di feedback positivo³²: 7 (*va bene*)
- Segnali di incomprensione: 1 (*come?*)
- Segnali di modalizzazione: 2 (*certo, decisamente*)

Anche nel sotto-corpus con iniziativa la strategia più usata per la riparazione di un errore è la ripetizione, ma, a differenza del sotto-corpus senza iniziativa in cui è quasi sempre ripetuto completamente il turno, incontriamo un maggior numero di variazioni:

³² Tra i segnali di feedback non sono state considerate le profrasi *sì* e *no*, normalmente preferite.



Graf. 5: Distribuzione delle ripetizioni/riformulazioni nei due sotto-corpus

In caso di situazione di errore possono attivarsi altre strategie, molto rare nel *corpus* senza iniziativa; infatti, in quasi la metà dei dialoghi sono presenti riformulazioni lessicali, che segnalano una maggiore propensione ad andare incontro al sistema e aiutarlo nella comprensione:

*WOM: vorrei / un treno / per il ritorno //

*MAC: non ho capito // può chiedere un' altra connessione / selezionare altri treni / o uscire // posso fare altro per lei ?

*WOM: sì // [///] voglio selezionare un altro [///] un' altra selezione / un altro percorso //

In quattro dialoghi, inoltre, è presente la ripetizione dialogica, completamente assente dal sotto-corpus senza iniziativa:

*MAC: vuole partire verso le ore venti e **usare un treno intercity** ?

*WOM: no // vorrei partire il mattino presto / e **usare sempre un treno intercity** //

Non è stato incluso tra le ripetizioni dialogiche il fenomeno dell'*alignment* (la tendenza a riusare forme, parole e costrutti dell'interlocutore), ampiamente presente nel sotto-corpus con iniziativa, soprattutto nei turni di correzione dopo una richiesta di conferma:

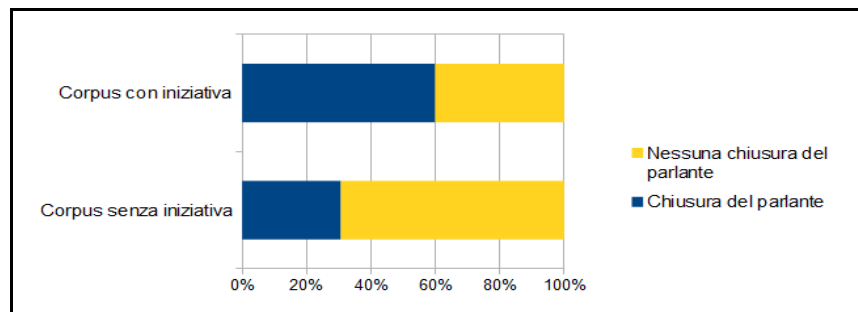
*MAC: dica la stazione di arrivo //

*WOM: **Roma** //

*MAC: **vuole arrivare a** Verona Porta Nuova ?

*WOM: no // **voglio arrivare a** Roma //

Infine, le conversazioni, oltre a concludersi spesso su iniziativa del parlante (come mostrato in 4.4.6), più raramente finiscono senza alcun segnale linguistico di chiusura, a differenza del sotto-*corpus* senza iniziativa, in cui più frequentemente si concludono con il soggetto che attacca il telefono:



Graf. 6: Distribuzione delle chiusure di conversazione nei due sotto-*corpus*

4.7 Conclusioni

Abbiamo visto come i due sotto-*corpus*, differenti per la presenza o l'assenza di turni di iniziativa del parlante, siano distanti sia sul piano linguistico sia sul piano pragmatico. Il corpus senza iniziativa presenta una scarsa varietà di elementi morfosintattici e costruzioni intrafrasali, tanto che più della metà dei dialoghi presentano solo turni con dati rilevanti per il sistema, senza alcuna costruzione; in questa tendenza a non intervenire se non per rispondere al sistema si innesta l'assenza di costruzioni macrosintattiche, l'assenza di segnali discorsivi e l'uso della ripetizione completa come unica strategia di riparazione degli errori: i parlanti del sotto-*corpus* senza iniziativa non sembrano variare quasi mai strategia in funzione del sistema, trattato come un'interfaccia in cui inserire dati piuttosto che un interlocutore.

Al contrario, il sotto-*corpus* di dialoghi con iniziativa presenta una *facies*

linguistica più eterogenea: gli elementi grammaticali sono presenti in maggior numero e varietà, così come formulazioni complete delle richieste, costruzioni macrosintattiche e segnali discorsivi. Il parlante sembra più propenso a variare ciò che ha detto per provare ad andare incontro al sistema: spesso adatta la richiesta alla formulazione dell'output; le ripetizioni in più della metà dei casi presentano una variazione; sono presenti, infine, ripetizioni dialogiche e riformulazioni lessicali, assenti dal corpus senza iniziativa. Il sotto-corpus con iniziativa mostra un comportamento tendenzialmente più attivo, ma è costituzionalmente più eterogeneo, non essendo stato possibile, date le piccole dimensioni, l'uso di filtri più sottili in funzione del diverso grado di iniziativa. Ciononostante, riprendendo i due poli di interazione con un sistema di dialogo (cfr. 3.4.3), possiamo far corrispondere i dialoghi senza iniziativa al polo dei parlanti che usano il “sistema come interfaccia”: a questo tipo di comportamento può ricondursi un *facies* linguistica scarna e omogenea, nonché una scarsissima varietà di strategie pragmatiche. Nel sotto-corpus con maggiore intervento da parte dell'utente si riscontrano invece una maggiore varietà di forme e strategie, più vicine al polo del “sistema come interlocutore”. Abbiamo visto come la gestione dell'iniziativa possa determinare gran parte della struttura dell'interazione e possa essere correlata alla varietà di forme e strategie usate dal parlante. In ottica applicativa, l'identificazione di almeno due profili basati sull'iniziativa può consentire al sistema di variare le proprie caratteristiche a seconda del parlante: ad esempio, la possibilità di passare dall'iniziativa mista all'iniziativa singola (architettura generalmente più precisa e robusta) potrebbe migliorare la precisione del sistema con i parlanti che usano il sistema come una semplice interfaccia. Naturalmente, la possibilità di identificare automaticamente il comportamento dell'utente potrà essere possibile solo con ulteriori studi sulla gestione dell'iniziativa nei dialoghi pratici e sulle eventuali correlazioni linguistiche e pragmatiche.

5. Il sistema simulato per la raccolta del *corpus* Speaky

5.1 Introduzione

Nella progettazione di un sistema di dialogo va trovato un compromesso tra l'elasticità e la robustezza: un sistema che lascia molta libertà all'utente può essere molto usabile poiché non costringe l'utente a dover apprendere alcunché del funzionamento e a svolgere ogni attività in linguaggio naturale; d'altra parte, minore è il controllo su quello che può dire l'utente in ogni stato della conversazione, maggiore è la probabilità che un sistema incorra in errate acquisizioni.

Tale compromesso va trovato anche tra le possibilità tecnologiche attuali e le esigenze che migliorano l'esperienza di un utente con un'interfaccia vocale. Queste ultime non coincidono necessariamente con l'uso quotidiano del linguaggio, sia per lo scopo della conversazione (principalmente focalizzata sullo svolgimento di un'attività), sia per la rilevanza di avere un agente non-umano come interlocutore (cfr. capitolo 3). Per questo, la raccolta e l'analisi dei dati linguistici, utili per costruire l'architettura del sistema e per scrivere le grammatiche di comprensione dei singoli stati di dialogo, necessitano di un compromesso tra il linguaggio naturale (cfr. 3.4) e le costrizioni imposte dall'interfaccia e dalle attività che il sistema deve svolgere.

I *corpora* di parlato uomo-uomo sono poco utili per la progettazione di un sistema di dialogo, presentando funzioni e usi normalmente esorbitanti rispetto all'interazione con un sistema di dialogo; inoltre, l'obiettivo di avvicinare l'interfaccia all'utente, eliminando il più possibile le regole di interazione da imparare e le costrizioni imposte dal sistema, non può prevedere una progettazione che prescindano dall'utente. Riprendendo le parole del manuale di annotazione di dialoghi per interfacce vocali (elaborato nell'ambito del progetto SI-TAL), per sviluppare un sistema di dialogo ci si trova nella particolare situazione in cui «occorrono i dati per sviluppare il sistema e, contemporaneamente, occorre il sistema per poter raccogliere i dati» (SI-TAL: 11).

La raccolta dei dati di interazione tra un utente e un sistema di dialogo, necessari alla ricerca e alla progettazione, può essere condotta simulando il sistema

di dialogo: per questo è stata elaborata una tecnica di elicitazione chiamata Wizard-of-Oz³³. La simulazione consiste nel far interagire una persona con una macchina “finta”, impersonata dallo sperimentatore (chiamato *wizard*), senza che il primo ne sia a conoscenza.

5.2 La tecnica del Mago di Oz³⁴

Il wizard deve sottostare ad alcune costrizioni affinché il suo comportamento somigli a quello del futuro sistema e affinché il soggetto sia convinto di interagire con un agente artificiale e non con un essere umano. Le precondizioni per ottenere dati utili sono tre (Fraser – Gilbert 1991: 82):

1) Per il wizard deve essere possibile simulare il sistema e con esso le attività da svolgere: se le applicazioni di cui il sistema costituirà l'interfaccia sono disponibili, il wizard dovrà usare tali applicazioni; nel caso le applicazioni non siano disponibili, deve essere possibile simularle. Ad esempio, un programma di calcolo non può essere simulato, essendo per un essere umano praticamente impossibile eseguire operazioni di calcolo velocemente quanto un computer; al contrario, se l'applicazione prevede delle operazioni facilmente eseguibili da un essere umano (ad esempio, riconoscere il contenuto di un'immagine), allora può essere simulata anch'essa.

2) Il wizard deve simulare il comportamento del futuro sistema: l'interazione con il sistema simulato deve essere condotta seguendo l'architettura del futuro sistema, per questo il wizard deve seguire un protocollo di comportamento affinché i dati siano utili per ciò che si intende sviluppare.

3) Il sistema simulato deve essere convincente: la condizione più importante è che il soggetto, per tutta l'interazione, creda di interagire con un agente artificiale; il realismo dell'interazione si ottiene se vengono rispettati i due punti precedenti, e se

³³ Il nome occorre per la prima volta in Kelley 1984, ed è derivato dal celebre romanzo di L. F. Baum). L'idea della simulazione risale agli anni '70, ed è precedente ai primi studi sulle differenze tra il dialogo uomo-uomo e il dialogo uomo-macchine. Per una storia dell'origine della simulazione cfr. Wei-Haas et al. 1985

³⁴ Per maggiori approfondimenti sulla tecnica del Mago di Oz e sui parametri da usare per la simulazione: Kelley 1984, Fraser – Gilbert 1991, Dahlbäck et al. 1993, Green et al. 2004, Steinfeld et al. 2009. Si veda inoltre l'analisi di Riek 2012, nell'ambito della interazione uomo-robot, sull'uso che viene fatto della tecnica del Mago di Oz.

prima e durante l'interazione il wizard rimane nascosto all'utente.

5.3 Variabili

Nei limiti imposti da queste precondizioni, lo sperimentatore deve tenere conto di alcune variabili, che replicano il contesto di interazione in cui il sistema si troverà a operare, e deve definire delle regole per il comportamento del wizard, in relazione al sistema che si intende progettare.

1. Ambito. Lo sperimentatore deve scegliere quali parti del sistema di dialogo simulare (ad es. il modulo di comprensione, il gestore del dialogo, ecc.). Nel caso siano disponibili (e facilmente integrabili nella simulazione) alcuni moduli, questi vanno usati nella simulazione.

2. Compito. Per coprire il più possibile le interazioni tra un utente e il futuro sistema di dialogo, vanno definiti i compiti che il soggetto deve svolgere, affini ai compiti che verranno svolti con il sistema reale. Al fine di rendere più realistica la simulazione, i compiti devono coinvolgere il più possibile le attività che il sistema dovrà svolgere e devono prevedere una soluzione che può essere raggiunta tramite più strade (Dahlbäck et al. 1993).

3. Soggetto. I soggetti devono essere rappresentativi del target di utenza finale del sistema: la diversa capacità, a seconda del soggetto, di comprendere gli enunciati prodotti dal wizard può condizionare la struttura del dialogo. Anche la produzione linguistica dei soggetti può variare (ad es. può esserci un forte influsso del dialetto o la presenza di forme non-standard), il wizard in questo caso deve scegliere quali forme accettare (cfr. punto 4). Inoltre, il soggetto può essere più o meno esperto del dominio di applicazione del sistema, e del sistema stesso (come già visto, l'acquisizione di esperienza del sistema cambia il comportamento linguistico del parlante). Infine, per una maggiore sicurezza sui dati acquisiti, Fraser – Gilbert 1991 consigliano un questionario finale in cui si accerta che la simulazione sia riuscita e che il parlante creda davvero di aver parlato con una macchina.

4. Wizard. Il ruolo del “mago” è naturalmente quello più delicato, le scelte legate al suo comportamento sono determinanti per la realizzazione del *corpus*. In funzione delle possibilità del sistema che verrà prodotto, vanno stabilite la capacità di comprendere gli enunciati del soggetto e la forma con cui generare gli enunciati del

sistema. Per quanto riguarda la capacità di comprensione, la raccolta può essere divisa in due fasi: nella prima vengono poste poche limitazioni linguistiche per la comprensione e la produzione, nella seconda vengono introdotte limitazioni su tutti i livelli linguistici (anche in base ai risultati della prima fase, vedi *infra*). Va inoltre simulato un certo numero di finti errori per studiare le strategie di correzione dell'utente. Dal punto di vista della produzione il wizard deve comportarsi in base al modulo di generazione del linguaggio usato nel sistema (più o meno rigidamente, a seconda della fase di raccolta). Da un punto di vista logistico il wizard deve prepararsi sul dominio di applicazione, sulle capacità del sistema e sugli strumenti che avrà a disposizione per poter accedere alle informazioni velocemente.

5. *Canale di comunicazione.* Il wizard deve comunicare con il soggetto tramite un sistema di sintesi vocale o con un vocoder che renda “meccanica” la sua voce, mentre la voce del soggetto può essere lasciata senza particolari condizionamenti (se si è scelto di introdurre gli errori “manualmente”, cfr. punto 4); tuttavia può essere applicato un disturbo alla voce del soggetto per ridurre la capacità di comprensione del wizard (aumentando però di molto il lavoro). Un'altra soluzione è interporre tra il soggetto e il wizard un riconoscitore vocale (solo nel caso sia lo stesso del futuro sistema di dialogo). Va inoltre definito se il segnale può passare contemporaneamente in entrambe le direzioni, stabilendo quindi, in funzione del futuro sistema, se permettere o no le sovrapposizioni.

5.4 Fasi della raccolta

La raccolta può essere suddiviso in due o più fasi (di raccolta).³⁵

- Fase pre-sperimentale: viene definito il dominio di applicazione, su cui deve esercitarsi il wizard. Vanno inoltre definiti gli scenari entro cui interagiranno i soggetti e i compiti che dovranno svolgere. Alcuni esperimenti pilota devono essere fatti per verificare se la progettazione degli scenari è idonea.

³⁵ Guyomard – Siroux 1987 hanno diviso la raccolta in due fasi: nella prima hanno raccolto dati sulla base sia di un modello di dialogo a iniziativa del sistema (e quindi del wizard) sia di un modello a iniziativa mista senza restrizioni, nella seconda hanno aggiunto restrizioni. Una suddivisione diversa è operata da Morel 1986: dialogo senza restrizioni nella prima fase, dialogo con restrizioni nella produzione linguistica del wizard (nessuna ellissi, nessuna anafora, nessuna domanda di conferma, nessuna ricapitolazione delle informazioni acquisite) nella seconda fase, dialogo con restrizioni anche nella comprensione del wizard (incomprensione per ellissi, anafore o sintassi complessa).

- Prima fase: serve alla definizione iniziale di lessico, grammatica e modello di dialogo; vanno quindi introdotte poche limitazioni linguistiche (il wizard mostrerà quindi di comprendere quasi tutto e parlerà liberamente).

- Seconda fase (ed eventualmente fasi successive): sulla base dei primi dati raccolti vanno aggiunte delle limitazioni, sia nella produzione sia nella comprensione del wizard (per esempio incomprensione per le anfore con antecedenti troppo lontani). Kelley 1984 propone di aggiungere via via, nelle fasi successive, componenti già sviluppate del sistema di dialogo (per esempio il riconoscitore vocale): in questo caso si parla di *bionic wizard*.

5.5 Critiche e limiti

Nonostante la tecnica del Mago di Oz sia attualmente il miglior metodo in circolazione per la progettazione di interfacce *user-oriented*, non mancano delle critiche e dei punti di debolezza di cui bisogna tenere conto. La critica principale riguarda la mancanza di realismo: per quanto si possa progettare una buona simulazione (generalmente, la maggior parte dei soggetti intervistati a fine sperimentazione non ha dubbi sul fatto che abbia interagito con una macchina), è stato fatto notare (Ogden 1988, Wiren et al. 2007) che i soggetti chiamati per un esperimento (e che quindi simulano a loro volta un ruolo), non hanno le stesse motivazioni e quindi lo stesso comportamento di utenti veri, realmente interessati allo svolgimento di un'attività³⁶. Per Dahlbäck et al. 1993 la critica ha senso solo per determinati aspetti controllabili dal soggetto, ma non per quelli non controllabili (come la maggior parte delle strutture linguistiche):

[relativamente alla simulazione di un sistema di prenotazioni di vacanze] «[...] if a user is just pretending to buy a holiday trip to Greece, she might not probe the alternative to the extent that she would if she were in fact to buy it, simply because the goal of finishing the task within a limited time takes precedence.

³⁶ Per ovviare alla mancanza di realismo e di reali motivazioni da parte dei soggetti coinvolti nella sperimentazione (rispetto a reali utenti), Wiren et al. 2007 hanno sperimentato la tecnica del Mago di Oz all'interno di un servizio telefonico realmente esistente: usando degli operatori telefonici, istruendoli come wizard e raccogliendo dati da telefonate reali. Con questa tecnica si risolvono alcuni problemi: le motivazioni sono reali, quindi più realistiche; la definizione degli scenari è superflua; i wizard non hanno bisogno di preparazione sul dominio, essendo già operatori; i wizard sanno interagire perfettamente con gli utenti. Naturalmente questa tecnica può essere usata principalmente per servizi telefonici, che possono contare su un grande flusso di dati.

But it does not seem likely that the latter fact will affect the use of pronouns in a specific utterance, or the knowledge about charter holidays and Greek geography, that is implicitly used in interpreting and formulating specific utterances» (p. 23).

Un altro limite della tecnica del Mago di Oz è l'impossibilità di poter simulare qualsiasi modulo del sistema di dialogo o qualsiasi compito, come ad esempio calcoli complessi alla stessa velocità con cui la eseguirebbe una macchina (Meier 2011).

Un'ultima critica è di ordine etico: è giusto "ingannare" delle persone a scopi scientifici? Il problema è molto sentito in alcuni stati: in Svezia, ad esempio, chiunque riceva dei fondi governativi per degli esperimenti è obbligato a far sapere tutto l'esperimento ai soggetti coinvolti. Dahlbäck et al. 1993 ricordano, però, che l'unico modo per raccogliere dei dati non viziati è non far sapere nulla al soggetto chiamato per la raccolta, e che, a esperimento finito, nessun soggetto si è lamentato o ha mostrato disappunto per essere stato vittima di uno "scherzo" a scopi scientifici.

5.6 La raccolta del *corpus* Speaky

5.6.1 Il progetto Speaky

La Fondazione Ugo Bordoni, nell'ambito della cui attività ho condotto la ricerca, partecipa al progetto Speaky nella progettazione del sistema di dialogo automatico (VUI, Voice User Interface), per aiutare alcune tipologie di persone (anziani e non vedenti) nella gestione della casa (se integrato in un impianto domotico), nell'uso di strumenti multimediali (come televisione, stereo, ecc.), per l'assistenza sanitaria e potenzialmente integrabile su servizi esterni (pagamenti, prenotazione biglietti, ecc.).

Il sistema di dialogo sarà integrato nella periferica hardware "Speaky Acutattile": un dispositivo di input/output *mouse-like*, che facilita l'interazione con ipovedenti e non vedenti. Si tratta di un telecomando/telefono a cui si possono dare sia comandi vocali sia comandi fisici, che vengono trasmessi in radiofrequenza al box-pc. A seconda del suo specifico profilo, l'utente avrà a disposizione uno o più microfoni senza fili. Il dispositivo funzionerà anche come touchpad per interfacce grafiche, essendo dotato di un ditale/magnete con cui spostare il puntatore sullo

schermo del pc. Quando il puntatore sullo schermo incontra un oggetto cliccabile, il ditale vibra e il sistema segnala all'utente, tramite la sintesi vocale, il nome dell'oggetto su cui si è posizionato. Questo dispositivo è anche in grado di aiutare l'utente non vedente nella ricerca degli elementi cliccabili sullo schermo: una microforza generata sul ditale da elettrocalamite laterali, conduce il ditale sull'elemento cliccabile più vicino lungo la direzione di spostamento iniziale. Collegato al box-pc vi sarà inoltre un sistema di sensori e attuatori domotici per l'interazione con l'ambiente domestico, per la gestione e il controllo di impianti elettrici, dell'illuminazione, del riscaldamento, ecc. Infine, vi sarà un sistema di monitoraggio posturale: il sistema seguirà gli utenti in casa e, monitorandone la postura, potrà segnalare le situazioni critiche (ad es. se la persona monitorata cade per terra). Nel sistema sono inoltre presenti dei sensori biometrici per la misurazione dei parametri vitali per la telemedicina e teleassistenza.

5.6.2 L'architettura dell'interfaccia vocale

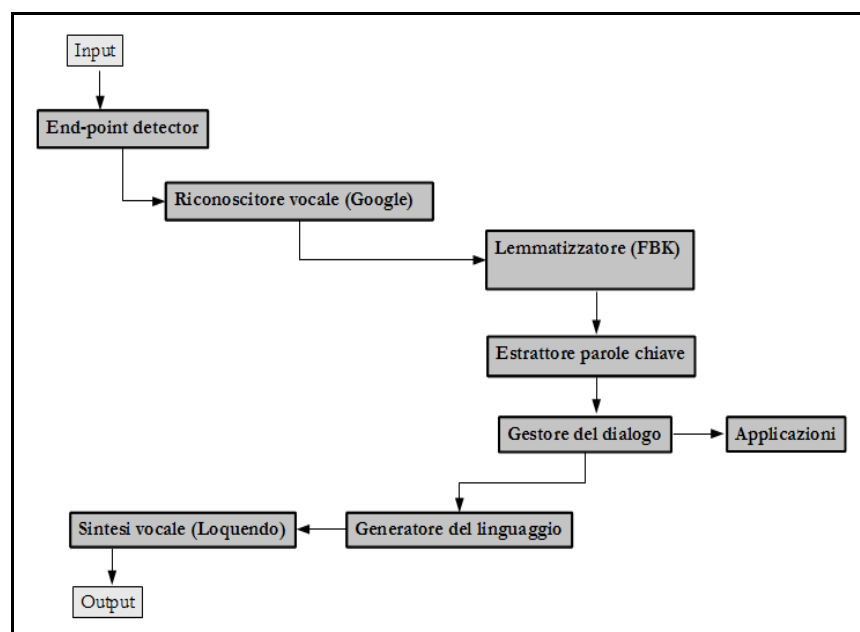


Fig. 3: Architettura del sistema di dialogo di Speaky Acutattile

L'interfaccia vocale di Speaky si basa su un'architettura *frame-based* (cfr. 2.4.3) ed è composta dai moduli raffigurati nella fig. 3, in parte sviluppati contestualmente al sistema di dialogo, in parte già commercializzati come

applicazioni. Il canale comunicativo tra parlante e sistema si apre tramite una parola chiave (“*speaky*”) individuata dal modulo di *end-point detection*, che durante l'interazione gestisce i turni di parola tra parlante e sistema; gli input vengono trascritti per intero dal riconoscitore vocale e successivamente lemmatizzati; dalla stringa lemmatizzata vengono quindi estratte le parole chiave: sulla base delle parole chiave usate il *dialogue manager* definisce lo stato del dialogo, a cui corrisponde sia l'attività da svolgere (in caso di completezza dell'input) sia l'output da inviare al parlante.

- *End-Point Detector*: il sistema rileva la presenza di parlato rivolto alla macchina tramite il SNR (Signal-to-Noise Ratio, cfr. 6.3.3.4) su microfinestre temporali di 50ms; il rilevamento parte 500ms dopo l'output del sistema. Anche il segmento di silenzio per individuare la fine del turno del parlante è di 500ms (cfr. 6.3.3.4). Non essendo concesso il *barge-in*, l'utente non può sovrapporsi e interrompere il sistema ma deve aspettare che finisca il turno.

- *Riconoscitore vocale*: il riconoscitore Google Speech trascrive per intero il segnale acustico; operando unicamente online, in caso di assenza di connessione, può essere sostituito dal riconoscitore vocale Cedat85.

- *Comprensione*: il modulo di comprensione semantica è composto da un lemmatizzatore (sviluppato dalla Fondazione Bruno Kessler), che si occupa di rimandare le parole usate dall'utente ai rispettivi lemmi, e da un estrattore di parole chiave: una volta lemmatizzata, la stringa di testo viene filtrata dall'estrattore di parole chiave sulla base di classi semantiche scritte a mano: al momento, le classi semantiche sono composte solo da lemmi di azioni e oggetti rilevanti per le attività del sistema (mentre deve essere ancora integrata la comprensione pragmatica). Le grammatiche di comprensione (ovvero, le classi semantiche) sono determinate dallo stato del dialogo (ad esempio, se l'utente sta svolgendo un'attività relativa alla gestione della casa, la classe semantica delle medicine è bloccata).

- *Gestore del dialogo*: l'acquisizione delle parole chiave determina il passaggio a un determinato stato del dialogo. Nello stato iniziale sono attive tutte le classi di parole chiave: se dal successivo input del parlante viene estratto un numero di parole chiave

insufficiente per lo svolgimento dell'attività, lo stato di arrivo comporta la richiesta delle variabili mancanti; se non viene individuata alcuna parola chiave, il sistema ne chiede la ripetizione e lo stato rimane invariato; se il numero di parole chiave è sufficiente allo svolgimento di un'attività, lo stato di arrivo comporta anche lo svolgimento dell'attività, e il dialogo torna automaticamente allo stato iniziale. Si veda l'esempio di uno stato con due parole chiave necessarie:

Numero di parole chiave atteso	Stato del dialogo (0 = parola chiave mancante; 1 = parola chiave estratta)	Output
2-key	(0;0)	Puoi dirmi data e orario?
1-key	(1;0)	Puoi dirmi l'orario?
1-key	(0;1)	Puoi dirmi la data?
0-key	(1;1)	Va bene, ho prenotato per il (data;orario).

Tab. 5: Modulo di comprensione di Speaky

- *Generazione del linguaggio e sintesi vocale*: a ogni stato corrisponde anche un determinato output, composto da una parte fissa e da un template variabile a seconda delle variabili fornite dall'utente (come l'ultimo output della tabella precedente). L'output viene quindi sintetizzato in tempo reale dal modulo di sintesi vocale, anch'esso prodotto da Google.

5.6.3 La raccolta del *corpus* Speaky

5.6.3.1 Ambiente e situazione extralinguistica

La principale condizione da rispettare affinché l'esperimento del Mago di Oz sia utile è che il parlante non sappia di interagire con un sistema simulato. La raccolta condotta a Roma è stata allestita in un laboratorio del Ministero dello Sviluppo Economico, composto da due stanze separate da uno specchio riflettente da un solo lato (quello dell'utente). Nelle raccolte di Palermo, Torino e Padova il wizard operava in remoto, collegato al computer con il software TeamViewer. In tutte le sessioni ogni soggetto sedeva a un tavolo, con davanti la lista dei compiti da svolgere e accanto uno degli sperimentatori che svolgeva il ruolo di accoglienza e assistenza

per potenziali problemi.³⁷ Il soggetto aveva davanti a sé uno schermo (a distanza di 3 metri circa) su cui erano trasmessi un avatar umanoide parlante (Mary, prodotto dall'Istituto di Scienze Cognitive di Padova) e un ambiente grafico legato al dominio e al compito da svolgere, come ad esempio menu di prodotti o trasmissioni televisive.

5.6.3.2 Soggetti

La scelta dei soggetti per l'esperimento è stata fatta sulla base delle tipologie di utenza a cui è rivolto Speaky: anziani (65-80 anni) e non vedenti. Complessivamente sono state raccolte le interazioni con 84 utenti, così suddivisi:

Roma: 19 anziani – 6 non vedenti

Torino: 9 anziani – 10 non vedenti

Padova: 10 anziani – 10 non vedenti

Palermo: 10 anziani – 10 non vedenti

Il *corpus* analizzato nella tesi è composto dalle interazioni raccolte a Roma e a Torino, in quanto le sessioni di Padova e Palermo sono state condotte in un secondo momento, ad analisi già iniziata.

5.6.3.3 Scenari

Così come i soggetti chiamati a interagire con il wizard devono essere della stessa tipologia a cui è rivolto il sistema di dialogo, allo stesso modo gli scenari devono simulare (e replicare) i contesti entro cui il sistema sarà usato, e quindi gli strumenti e i servizi che dovrà gestire. Con scenari si intendono quindi i contesti di interazione simulati (casa domotica, mediacenter, ecc.) in cui i soggetti e i wizard interagiscono allo scopo di risolvere determinati compiti. Poiché il sistema di dialogo di Speaky è un'interfaccia non legata a precisi dispositivi (ad esempio a una determinata casa domotica con le proprie caratteristiche tecniche), non ci si è potuti attenere a una lista di funzioni già definite; abbiamo quindi assegnato, a ognuno degli scenari, un numero di funzioni base, a cui ci siamo attenuti per la scrittura dei

³⁷ Questa scelta obbligata dalle necessità dei soggetti rappresentativi delle classi di utenza condiziona naturalmente il parlato al sistema; d'altra parte, si è rivelata inaspettatamente utile per ricavare di tanto in tanto commenti metalinguistici dei soggetti durante l'interazione con il sistema.

compiti. Gli strumenti e i servizi sono stati quindi raggruppati in quattro macro-scenari:

- Domotica: scenario relativo alla gestione della casa (riscaldamento, illuminazione, finestre, tapparelle);
- Assistenza sanitaria: gestione del proprio profilo sanitario (database analisi, farmaci, ecc.);
- Mediacenter: uso della televisione, uso dello stereo, pay-per-view;
- Servizi esterni: prenotazione taxi, acquisto e prenotazione biglietti (treno, cinema, teatro).

Gli scenari che prevedono normalmente l'interazione con un terminale (intrattenimento, assistenza sanitaria, servizi esterni) sono facilmente replicabili: se, per esempio, l'utente chiede informazioni sul palinsesto televisivo, questo viene visualizzato sullo schermo. Lo scenario domotico, invece, avrebbe richiesto un'effettiva casa domotica entro cui ambientare la simulazione: non avendo avuto a disposizione questa, l'intera ambientazione è stata simulata con un ambiente virtuale interattivo, visualizzato sempre sullo schermo, in cui il soggetto ha potuto svolgere le attività (naturalmente per via vocale) e vedere i risultati delle proprie azioni (apertura/chiusura di una finestra, regolazione della luminosità, ecc.).



Fig. 4: Ambiente domotico virtuale

5.6.3.4 Compiti

I compiti (o *task*) sono le attività che il soggetto deve svolgere tramite l'interfaccia vocale simulata e sono scritti sulla base delle funzioni che il sistema dovrà svolgere.

Sono stati stilati in totale 48 compiti: ogni soggetto ha dovuto svolgere 19-20 compiti diversi³⁸, composti ognuno da 2-3 sotto-compiti (o *subtask*) connessi tra loro. I compiti sono stati scritti coerentemente al modulo di comprensione del futuro sistema Speaky, secondo un modello *frame-and-slot* (cfr. 2.3.2): ogni sottocompito prevede perciò un certo numero di variabili da fornire al sistema (di cui alcune obbligatorie e altre facoltative) per il completamento dell'attività.

Descrizione	Descrizione estesa	Sotto-compiti
Impostare gli orari per l'assunzione di alcuni medicinali e scegliere la modalità di avviso.	Il soggetto deve dare il nome del medicinale, la quantità, l'orario d'assunzione ed eventualmente il giorno; successivamente il soggetto deve definire la modalità d'avviso (SMS, sveglia ed eventuale avviso sonoro, avviso vocale del sistema) e provarla subito.	Medicinali: (S1) nome, (S2) quantità, (S3) orario, (S4) giorni della settimana. Promemoria: (S1) tipologia d'avviso (SMS/avviso vocale del sistema/avviso sonoro), (S2) provare promemoria.
Consultare la lista di analisi richieste dal medico di base e prenotarle.	Dopo l'ultima visita, il medico ha inviato via e-mail (o inserito nel database del paziente) una lista di analisi: il soggetto deve verificare se è stata inserita, consultarla e prenotare ogni analisi, definendo insieme al sistema, in base alla disponibilità, data e orario; il soggetto può anche decidere, dopo richiesta del sistema, di	Lista analisi: (S1) consultare lista, (S2) prenotazione analisi, (S2-1) orario, (S2-2) data, (S2-3) luogo. Pagamento ticket: (S1) info costo, (S2) info saldo carta di credito, (S3)

³⁸ L'oscillazione è dovuta alla presenza o meno, nelle sessioni, dell'ultimo compito da svolgere, ovvero un'interazione diretta con lo scenario domotico che consisteva nel regolare l'altezza delle tapparelle. Per limitazioni tecniche non è stato possibile usare lo scenario domotico interattivo nelle prime quattro sessioni a Roma e in tutte le sessioni condotte in remoto.

Descrizione	Descrizione estese	Sotto-compiti
	pagare subito i ticket.	pagamento.
Chiedere consiglio e ordinare un farmaco generico contro l'emicrania.	Il soggetto deve segnalare al sistema una qualche indisposizione fisica (emicrania, ma può essere personalizzato) e chiedere consiglio su un qualche farmaco da prendere; successivamente deve acquistarlo online, verificandone la compatibilità con altri farmaci che prende, il prezzo e il tempo di consegna.	Consiglio farmaco: (S1) dolore. Acquisto farmaco: (S1) compatibilità farmaci, (S2) info costo, (S3) info saldo carta di credito, (S4) info tempo di consegna, (S5) pagamento.
Prendere degli appuntamenti per 2 sedute di igiene dentale.	Prenotare 2 appuntamenti entro due mesi, definendo data e orario; chiedi inoltre di farti preparare la fattura per entrambe le sedute già al primo appuntamento per poterla subito consegnare all'assicurazione sanitaria per il rimborso.	Prenotazione: (S1) data, (S2) orario. Fattura: (S1) richiesta fattura.

Tab. 6: Struttura dei compiti da far svolgere ai soggetti

Come vedremo nel paragrafo 5.6.3.5.2 abbiamo scelto di usare degli output fissi preregistrati e abbiamo limitato l'uso della sintesi in tempo reale: per questo siamo stati obbligati a fornire al soggetto anche le variabili³⁹ per tutti quei compiti che avrebbero previsto un numero troppo grande di combinazioni, e conseguentemente di output.

Una delle variabili critiche nella raccolta di parlato semi-spontaneo è l'addestramento dei soggetti sui compiti da svolgere (cfr. Cerrato 2006): una descrizione dettagliata dei compiti permette di avere un alto controllo sulla produzione, affinché vengano completati tutti i sotto-obiettivi senza troppe interruzioni, ma condiziona il soggetto sia dal punto di vista lessicale, sia per quanto

³⁹ Ad esempio, nel compito che prevedeva la prenotazione di un biglietto del treno, sono state fornite anche la data e gli orari («Devi partire il ventisette aprile, di mattina, per Milano»), mentre il soggetto non aveva particolari indicazioni riguardo alla classe del treno o altre informazioni da chiedere.

riguarda lo svolgimento del compito; d'altra parte, l'uso di vignette o story-board per spiegare i compiti da svolgere condiziona in misura minore il soggetto, ma allo stesso tempo, comporta una maggiore difficoltà di previsione degli input e quindi un maggiore carico di lavoro per il wizard, che può portare la simulazione a non rispettare la prima preconditione della tecnica del Mago di Oz («per il wizard deve essere possibile simulare il sistema e con esso le attività da svolgere», cfr. 5.1).

Le istruzioni sono state impartite in due modalità: inizialmente è stata letta al soggetto una descrizione del compito da svolgere, per contestualizzare gli obiettivi da raggiungere; successivamente, durante il compito, il soggetto si è potuto appoggiare a una tabella riepilogativa con le variabili necessari per lo svolgimento del compito.

5.6.3.5 Simulazione del sistema

Il ruolo di wizard è stato svolto da 3 diversi soggetti. Per avere una condotta omogenea con ogni soggetto coinvolto nell'esperimento e per avere una rapida reazione agli input da parte del parlanti, il comportamento del wizard è stato indirizzato da alcuni parametri definiti a priori.

Sul lato della comprensione, il wizard ascoltava in cuffia il parlato dell'utente, senza avere particolari costrizioni: il parlante umano non è stato condizionato da limitazioni legate a uno specifico modulo di comprensione del linguaggio e quindi a una specifica forma con cui presentare la richiesta; sono state, tuttavia, rifiutate richieste semanticamente non attinenti al dominio (con semplici richieste di ripetizione, del tipo «Non ho capito, puoi ripetere?»). Per non avere interazioni prive di errori e non realistiche, le situazioni di errore sono state simulate, seguendo le indicazioni che saranno illustrate nel paragrafo 6.4.4.

Sul lato della produzione, invece, il wizard ha usato un set di output registrati, cercando di evitare l'uso del sistema di sintesi vocale in tempo reale, anch'esso possibile. Gli output sono stati organizzati, sulla base della struttura apertura-negoziazione-chiusura, in alberi di dialogo, così da permettere al wizard di sapere cosa rispondere a seconda degli input del soggetto e dello stato del dialogo.

5.6.3.5.1 Interfaccia del wizard

Il wizard ha operato tramite un'interfaccia grafica su cui selezionare gli output

da far dire al sistema. Ogni compito ha un suo set di output scritti e registrati con un sistema *text-to-speech* da caricare all'inizio del compito stesso; inoltre, era a disposizione un set di output generici da usare a seconda della necessità:

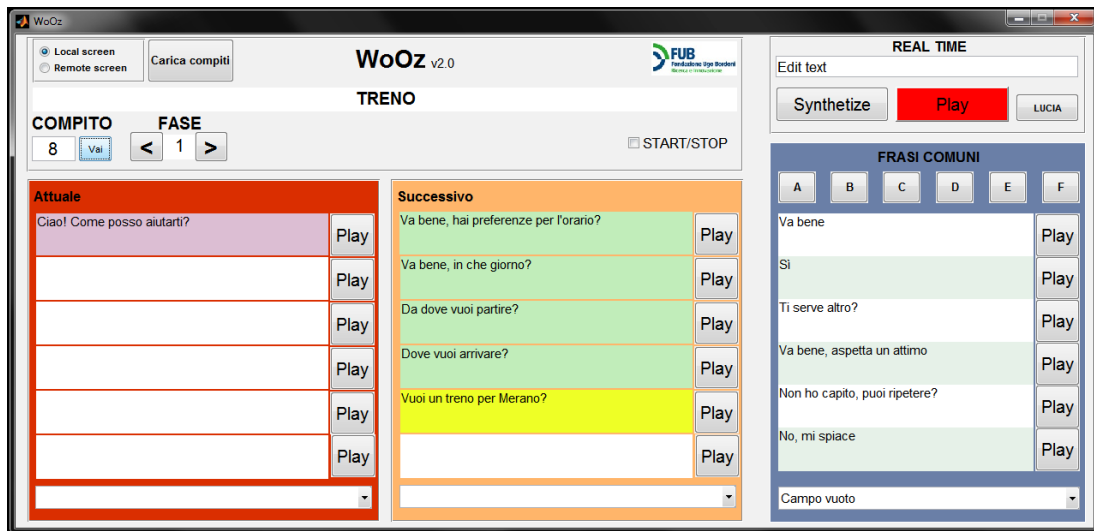


Fig. 5: Interfaccia del wizard

5.6.3.5.2. Definizione dei set di output e dell'albero di dialogo

Per ogni compito è stato preparato un set di output, da usare in consonanza con gli obiettivi del compito. Ogni set è diviso in sottocompiti, fase del dialogo e tipologia dell'output. Come illustrato nella tabella d'esempio, la prima colonna definisce il sottocompito, la seconda colonna, la fase del dialogo (1 = apertura, 2 = negoziazione, 3 = chiusura) mentre la terza il tipo di output (1 = apertura generica, 2 = apertura guidata, 3 = richiesta di completamento, 4 = richiesta di ripetizione, 5 = richiesta di conferma errata, 6 = feedback di completamento).

1	1	1	Ciao! Come posso aiutarti?
1	2	3	Va bene, hai preferenze per l'orario?
1	2	3	Va bene, in che giorno?
1	2	3	Da dove vuoi partire?
1	2	3	Dove vuoi arrivare?
1	2	5	Vuoi un treno per Merano?
			Il 27 aprile c'è un treno per Milano Centrale che parte da Roma Termini alle
1	3	6	10 e 40.
2	1	1	Ti serve altro?
2	1	2	Vuoi cercare anche il ritorno?
2	2	3	Va bene, hai preferenze per l'orario?
2	2	3	Va bene, in che giorno?

2	3	6	Per il ritorno ce n'è uno che parte alle 9 e 30 da Milano Centrale.
2	3	6	Per il ritorno ce n'è uno che parte il 28 aprile alle 9 e 30 da Milano Centrale.
3	1	1	Posso esserti ancora utile?
3	1	2	Vuoi altre informazioni?
3	3	3	Per quale classe?
3	3	3	Vuoi sapere la durata del viaggio d'andata o di quello di ritorno?
3	3	6	Roma - Milano: Il costo del biglietto di prima classe è 80 euro.
3	3	6	Roma - Milano: Il costo del biglietto di seconda classe è 45 euro.
3	3	6	Roma - Milano: La durata del viaggio è 3 ore e 45 minuti.
3	3	6	Milano - Roma: Il costo del biglietto di prima classe è 85 euro.
3	3	6	Milano - Roma: Il costo del biglietto di seconda classe è 48 euro.
3	3	6	Milano - Roma: La durata del viaggio è 3 ore e 35 minuti.
4	1	1	Ti serve altro?
			Se vuoi puoi procedere con l'acquisto del biglietto tramite carta di credito,
4	1	2	altrimenti puoi prenotare il biglietto.
4	1	2	Vuoi acquistarlo o prenotarlo?
4	2	4	Non riesco a comprendere, puoi ripetere?
4	2	3	Va bene, per l'acquisto dovresti dirmi quale carta hai e il suo numero.
			Va bene, per la prenotazione dovresti darmi il tuo nominativo e un numero
4	2	3	di cellulare.
			Hai acquistato un biglietto per Roma-Milano il 27 aprile alle 10 e 40, e un
4	3	6	biglietto di ritorno per il 28 aprile alle 9 e 30.
			Hai prenotato un biglietto per Roma-Milano il 27 aprile alle 10 e 40, e un
4	3	6	biglietto di ritorno per il 28 aprile alle 9 e 30.
4	3	6	Hai acquistato un biglietto per Roma-Milano il 27 aprile alle 10 e 40.
4	3	6	Hai prenotato un biglietto per Roma-Milano il 27 aprile alle 10 e 40.
4	3	6	Hai acquistato un biglietto per Milano-Roma il 28 aprile alle 9 e 30.
4	3	6	Hai prenotato un biglietto per Milano-Roma il 28 aprile alle 9 e 30.

Tab. 7: Classificazione degli output di un compito

Il set generico è invece uguale per ogni compito e comprende gli output il cui uso è esteso a ogni interazione, come i feedback di accordo, i saluti e le risposte alle richieste fuori dominio:

Va bene
Ciao
Sì
No, questo non posso farlo
Va bene, aspetta un attimo
Non ho capito, puoi ripetere?
Ti serve altro?
Mi spiace ma non so risponderti
No, mi spiace

Tab. 8: Output generici

5.6.3.5.3 Output di apertura del compito

Ogni dialogo è iniziato con un input di attivazione del parlante (cfr. 6.3.2.1), a cui è seguito un output del tipo “How may I help you?” (Gorin et al. 1997), con cui viene lasciata l'iniziativa al parlante per indicare l'attività da svolgere e, potenzialmente, per organizzarne la risoluzione in un solo turno. A seguito del completamento della prima attività, mancando un'eventuale apertura di quella successiva da parte dell'utente (entro tre secondi, cfr. 6.3.2.1), è spettato al wizard indirizzare il dialogo sul secondo sottocompito con una richiesta di apertura generica («Ti serve altro?»), o, in casi particolari, indirizzando l'apertura del compito⁴⁰. Dopo l'ultima attività del compito, pur non essendo prevista un'attività successiva, il wizard chiedeva comunque al soggetto se intendeva svolgere un'altra attività.

5.6.3.5.4 Feedback di completamento

Una volta che i dati necessari sono stati acquisiti dal Mago, il completamento dell'attività è segnalato da un feedback, che può essere, a seconda della natura del compito, un feedback che segnala il corretto svolgimento di un'attività o la presentazione delle informazioni richieste.

5.6.3.5.5 Fase di negoziazione

La fase di negoziazione è caratterizzata da diversi tipi di output: richieste di completamento, richieste di ripetizione, richieste di conferma.

Successivamente all'apertura del compito, l'iniziativa passa al parlante, il cui input può includere tutte le informazioni necessarie o, altrimenti, mancare di qualche variabile; in tal caso, l'iniziativa torna al wizard, che ha il compito di chiedere i dati mancanti con richieste di completamento predisposti per coprire ogni caso possibile di assenza di informazioni.

Alla fase di negoziazione appartiene anche la casistica di errori e di recupero degli input non compresi dal sistema. Avendo lasciato libertà al wizard per quel che riguarda la comprensione, per evitare che il *corpus* non contenesse dati riguardanti tali situazioni, sono state simulate richieste di ripetizione e richieste di conferma. Anche in questo caso è stato rispettato per gran parte un protocollo definito *a priori*

⁴⁰ Di norma si è preferito usare la richiesta di apertura generica, per avere maggiori dati sulle aperture di compiti diversi, vista la natura multi-task del sistema *Speaky*; tuttavia, a fronte di eccessive esitazioni nell'apertura (anche dopo la prima richiesta), abbiamo scelto di includere output focalizzati sull'attività successiva, per non interrompere l'interazione.

in cui gli errori sono posti in punti diversi dell'albero di dialogo, per poter analizzare le strategie di riformulazione e di correzione in relazione allo stato del dialogo e ai turni "incompresi" (cfr. 6.3). Ogni compito prevedeva l'uso di una richiesta di ripetizione («Non ho capito, puoi ripetere?», «Non riesco a comprendere, puoi ripetere?», «Non ho capito, potresti ridirmelo?»). Prevedeva inoltre una richiesta di conferma errata, scritta appositamente per il compito con una variabile errata, da usare coerentemente con le informazioni presenti nell'input del parlante.

5.6.3.5.6 Gestione dell'iniziativa

Per la gestione del dialogo abbiamo usato un modello a iniziativa mista naturale (cfr. 2.4.1) consentendo al parlante di non essere rigidamente legato alle richieste del sistema. Ad esempio, a fronte di una richiesta di conferma errata, il parlante non era limitato a rispondere negativamente al wizard e aspettare le indicazioni del sistema per la correzione, ma ha potuto correggere l'informazione direttamente nel turno successivo a quello del wizard (es. W: «Vuoi avere informazioni sui treni da Roma a Torino?» – U: «No, da Roma a Milano»); allo stesso modo, a fronte di una richiesta specifica su una determinata variabile, era consentito fornire più informazioni (naturalmente rilevanti) di quanto richiesto esplicitamente dal sistema (es. W: «Dove vuoi andare?» – U: «Da Roma a Milano il 27 aprile»).

Per quel che riguarda la gestione del turno, a parte la domanda iniziale del wizard, è stato consentito al parlante (in possesso di una descrizione esaustiva del compito) di prendere il turno (e l'iniziativa) subito dopo la chiusura di un sotto-compito; mancando questa, il wizard ha preso il turno con una domanda di apertura del secondo sotto-compito. Per limitazioni tecniche legate ai movimenti facciali dell'avatar, al soggetto non è stato consentito interrompere l'output del sistema con eventuali sovrapposizioni (il cosiddetto *barge-in*), il che può aver causato, come già detto, una minore presenza del fenomeno.

5.6.3.6 Durata delle sessioni.

Ogni soggetto ha sostenuto una sessione di venti compiti l'una. La durata media di ogni sessione è stata di 43 minuti. Il *corpus* analizzato nella tesi comprende 31 ore e mezza di parlato, mentre quello complessivo è composto da 60 ore di parlato.

6. Analisi del *corpus* Speaky

6.1 Introduzione

L'analisi del *corpus* Speaky è focalizzata su due aspetti dell'interazione uomo-macchina: la risoluzione degli errori e la gestione dell'iniziativa e del turno da parte degli utenti. Naturalmente il materiale del *corpus* può essere analizzato da altri punti vista e su diversi livelli linguistici: dalla formulazione delle richieste all'uso di alcune classi di segnali discorsivi, dall'allocuzione alla prosodia. La scelta di focalizzarsi su questi due aspetti è motivata da alcune ragioni: innanzitutto, entrambi gli argomenti hanno un interesse condiviso tra linguisti e ingegneri. La risoluzione degli errori è stata indagata sin dagli anni '90, per le ovvie ragioni legate agli scarsi risultati dei riconoscitori vocali (prima) e per il supporto che può dare a interfacce più elastiche e perciò meno robuste e potenzialmente problematiche; inoltre, il parlante, nell'interazione con un agente non umano di cui non conosce il funzionamento, è spinto a dover fare ipotesi sulle cause dell'errore e sul processo di comprensione dell'agente (Fischer 2003, 2006a, entrambi confluiti in Fischer 2006b), comportandosi di conseguenza nella riformulazione del proprio turno. La gestione del turno e dell'iniziativa ha invece una storia più recente: inizialmente, con architetture molto restrittive, cosa un parlante poteva dire (iniziativa) era definito dal sistema, e quando poteva dirlo (turno) era stabilito con un'azione fisica (*push-to-talk*) o con un segnale acustico; negli ultimi 15 anni la ricerca ha cominciato a concentrarsi sull'argomento per rendere la gestione più affine a quella dell'interazione umana e meno meccanica. All'interesse della comunità scientifica si lega naturalmente la più ampia presenza di lavori con cui confrontare metodi e risultati, per quanto, come abbiamo visto, spesso le ricerche sperimentali sono difficilmente confrontabili per via della diversità tra sistemi e necessitano, quindi, della piena esposizione delle meccaniche del sistema (cfr. capitolo 5). Proprio il contesto sperimentale obbliga di per sé a focalizzare l'attenzione su quegli aspetti delle interazioni che sono meno condizionati dalle istruzioni impartite ai soggetti. Il target principale del sistema, composto da anziani e non vedenti, ha reso necessaria sia la presenza di uno dei ricercatori nella stanza del soggetto che interagiva con il

computer, sia l'uso del foglio riepilogativo delle attività durante lo svolgimento di queste: questo ha naturalmente condizionato le scelte lessicali e la formulazione degli input. Per lo sviluppo del sistema tale condizionamento non costituisce un ostacolo: l'allargamento del dizionario e il miglioramento delle possibili situazioni nel singolo turno di dialogo andranno certamente considerati e implementate in una fase successiva del progetto, con dati ottenuti dall'uso del sistema reale in condizioni non sperimentali ma di uso comune. Al contrario, la risoluzione degli errori e la gestione del turno sono meno legate alle istruzioni ricevute e dipendono maggiormente dalle situazioni dialogiche con cui il soggetto si è dovuto confrontare nell'interazione con il sistema.

6.2 Caratteristiche del *corpus*

Come anticipato (cfr. 5.6.3) il *corpus* selezionato per la tesi è costituito dalle interazioni registrate nelle sessioni di Roma e Torino: in totale 44 parlanti con cui sono stati raccolti 860 dialoghi (corrispondenti ad altrettanti parlanti). Ogni parlante ha interagito con il sistema per svolgere tra i 19 e i 20 compiti, a seconda se fosse possibile simulare il sistema domotico (cfr. 5.6.3.2).

La lunghezza delle sessioni e del numero di turni per sessione è naturalmente diversa da parlante a parlante, a seconda di alcune variabili: innanzitutto, i compiti svolti, che non sono stati gli stessi per tutti; poi, le scelte compiute durante l'interazione hanno condizionato il numero dei turni, a seconda di come il parlante ha svolto il compito (sintetizzando tutte le variabili da fornire al sistema in un solo turno o preferendo insieme o diluite), e a seconda delle scelte che ha fatto durante lo svolgimento (ad esempio, chi ha provato a testare il sistema con richieste fuori dominio, o chi si è dilungato nella scelta di prodotti da acquistare); infine, il tempo impiegato nel rispondere al sistema (e il tempo che ha impiegato il sistema a rispondere, in parte variabile, cfr. 5.6.3.5.6). Di media ogni dialogo è durato poco più di 2 minuti e 21 turni, per un totale di 40 minuti e 407 turni di conversazione a sessione. La durata di ogni sessione è riportata nella tabella riepilogativa: ogni sessione è contrassegnata dalla tipologia del soggetto (AN per anziano e NV per non vedente) e da un numero progressivo.

Soggetto	Turni sessione	Minuti sessione	N° Dialoghi	Media turni per dialogo	Durata media per dialogo (min)	Città
An01	442	38	19	23,3	2,0	Roma
An02	464	42	19	24,4	2,2	Roma
An03	420	37	20	21,0	1,9	Roma
An04	435	43	20	21,8	2,2	Roma
An05	372	45	20	18,6	2,3	Roma
An06	459	55	20	23,0	2,8	Roma
An07	308	60	20	15,4	3,0	Roma
An08	314	55	20	15,7	2,8	Roma
An09	352	35	20	17,6	1,8	Roma
An10	357	39	20	17,9	2,0	Roma
An11	363	41	20	18,2	2,1	Roma
An12	372	38	20	18,6	1,9	Roma
An13	442	42	20	22,1	2,1	Roma
An14	428	48	20	21,4	2,4	Roma
An15	467	53	20	23,4	2,7	Roma
An16	453	52	20	22,7	2,6	Roma
An17	560	44	20	28,0	2,2	Roma
An18	481	47	20	24,1	2,4	Roma
An19	443	45	20	22,2	2,3	Roma
An20	413	44	19	21,7	2,3	Torino
An21	374	30	19	19,7	1,6	Torino
An22	362	32	19	19,1	1,7	Torino
An23	365	35	19	19,2	1,8	Torino
An24	410	38	19	21,6	2,0	Torino
An25	398	45	19	20,9	2,4	Torino
An26	384	30	19	20,2	1,6	Torino
An27	371	37	19	19,5	1,9	Torino
An28	474	53	19	24,9	2,8	Torino
Nv01	540	43	20	27,0	2,2	Roma
Nv02	558	50	20	27,9	2,5	Roma
Nv03	424	48	20	21,2	2,4	Roma
Nv04	360	45	20	18,0	2,3	Roma
Nv05	348	42	20	17,4	2,1	Roma
Nv06	460	50	20	23,0	2,5	Roma

Nv07	370	35	19	19,5	1,8	Torino
Nv08	375	40	19	19,7	2,1	Torino
Nv09	404	35	19	21,3	1,8	Torino
Nv10	403	35	19	21,2	1,8	Torino
Nv11	361	44	19	19,0	2,3	Torino
Nv12	341	46	19	17,9	2,4	Torino
Nv13	345	34	19	18,2	1,8	Torino
Nv14	315	30	19	16,6	1,6	Torino
Nv15	327	32	19	17,2	1,7	Torino
Nv16	476	50	20	23,8	2,5	Torino

Il database vocale è composto dalle registrazioni di tutte le sessioni, comprese le istruzioni iniziali e le eventuali interazioni tra soggetto e sperimentatore presente nella stanza; ogni sessione è stata registrata su più tracce, audio e video, sincronizzate:

- 2 Microsoft Kinect (WAV PCM, audio mono 24-bit/16 kHz)
- 1 Radiomicrofono Sennheiser
- 2 Videocamera ZOOM Q3HD (MPEG-4 AVC/H.264 720p, audio stereo 24-bit/48 kHz)
- Samsung Galaxy SII (WAV PCM, audio mono 16-bit/44.1 kHz)

Il corpus è disponibile anche in formato testuale, trascritto a partire dalla registrazione fatta tramite il radiomicrofono, usando gli altri file solo nei casi di dubbio su singoli turni. Al momento non sono stati presi in considerazione i dati video (che riprendono i movimenti e le espressioni del soggetto da due diverse angolazioni). Il corpus testuale è stato sincronizzato alle tracce audio tramite il software Transcriber 1.5.1. Visto l'allineamento del testo con i file audio, che consente di recuperare rapidamente segmenti di dialogo, la trascrizione è stata di tipo ortografico, organizzata per turni. Sono stati tuttavia annotati fenomeni dialogici tipici, come pause, pause piene, false partenze e sovrapposizioni.

6.2.1 Criteri della trascrizione

La rappresentazione scritta delle interazioni non è stata sovraccaricata, poiché l'allineamento con i file audio consente di recuperare rapidamente la parte registrata di un qualsiasi segmento di dialogo; per questo la trascrizione è di tipo ortografico, organizzata per turni. Tuttavia, oltre alla trascrizione di ciò che è stato detto durante l'interazione, sono state annotati fenomeni notevoli del parlato dialogico, sia interni al singolo turno, sia legati alla meccanica di gestione dei turni. Si illustrano le convenzioni usate per la trascrizione, sempre ricordando la possibilità di poter tornare sui singoli segmenti audio per annotazioni più precise, a seconda dell'obiettivo dell'indagine (ad esempio, un'analisi della gestione del turno implica una maggiore precisione sulla durata delle pause):

- la trascrizione del dialogo è organizzata per turni di conversazione; il turno del mago è segnalato da “WOZ”, quello del soggetto da “AN” (anziano) o “NV” (non vedente);
- tutte le trascrizioni sono in minuscolo;
- le pause vuote all'interno del proprio turno sono state trascritte con la “,” per le pause brevi (tra 0,5s e 1,5s), e con il “.” per le pause lunghe (più di 1,5s);
- le pause piene sono trascritte con il fonema relativo tra i segni minore e maggiore, ad esempio <aa>, <ee>, <mm>, ecc.;
- le pause piene corrispondenti ad allungamenti di vocale o consonante, iniziali o finali, sono stati trascritti accanto al fonema interessato, con la dicitura <vv> se vocali e <cc> se consonanti. Ad esempio, *televisione*<vv>, *un*<cc>, ecc.;
- fenomeni contestuali come parole interrotte, non finite, o semplici errori di pronuncia sono marcate con un asterisco finale, con lo scopo di distinguerle dal lessico reale, ad esempio: “voglio sapere le previsioni mete* meteo”;
- le parole inintelligibili sono trascritte con un doppio asterisco **;
- le false partenze e le riformulazioni interne al turno sono segnalate da uno slash tra l'enunciato interrotto e la ripartenza, ad esempio “va bene allora, ah potrei avere / potresti acquistarmelo, a nome di mario rossi”;
- altri fenomeni vocali non verbali interni al turno come colpi di tosse, risate, ecc. sono segnalati tra parentesi quadre con la relativa terminazione inglese, ad es.

[cough], [laugh] (in inglese, per distinguere dall'altro uso che viene fatto delle parentesi quadre nella trascrizione, vd. *infra*).

- tra parentesi quadre sono segnalate situazioni di *offtalk*, ovvero di parlato non rivolto alla macchina (come qualcosa detto tra sé e sé, o rivolto all'accompagnatore); sono state trascritte per intero le parti offtalk con commenti metalinguistici relativi all'interazione (ad esempio: “ok grazie [glielo devo rispondere grazie? però però lei lei non risponde. non dice prego, allora non glielo dico più]”);
- le sovrapposizioni sono segnalate dall'uso di due segni cancelletto # che delimitano la parte del turno che avviene in sovrapposizione; da un punto di vista spaziale, il turno viene “spezzato” graficamente se alla sovrapposizione corrisponde un tentativo di presa del turno:

Woz: non ho capito
Nv02: # ma* #
Woz: # puoi # ripetere?
Nv02: massimo

Nel caso di feedback e segnali di assenso, invece, il turno non viene spezzato, non essendo tentativi di presa del turno:

Woz: nei prossimi giorni ## è previsto un abbassamento della
temperatura fino a due gradi
An16: # sì #

6.2.2 Un esempio di dialogo con il sistema

An02: ciao speaky
Woz: ciao, come posso aiutarti?
An02: ho bisogno di prenotare, un biglietto per milano
Woz: non ho capito, puoi ripetere?
An02: ho bisogno di prenotare un biglietto sul treno. per milano
Woz: vuoi un treno per merano?
An02: milano, mi_la_no
Woz: va bene, hai preferenze per l'orario?

An02: sì le<ee> dunque, di mattina il ventisette aprile

Woz: il ventisette aprile c'è un treno per milano centrale che parte da roma termini alle dieci e quaranta

An02: va bene

Woz: vuoi cercare anche il ritorno?

An02: sì, il ventotto aprile sempre di mattina

Woz: per il ritorno ce n'è uno che parte alle nove e trenta da milano centrale

An02: va bene

Woz: vuoi altre informazioni?

An02: sì, il costo del biglietto in seconda classe

Woz: roma - milano: il costo del biglietto di seconda classe è quarantacinque euro

An02: # va* #

Woz # milano #- roma: il costo del biglietto di seconda classe è quarantotto euro

An02: va bene

Woz: ti serve altro?

An02: sì ho bisogno di sapere la durata del viaggio

Woz: vuoi sapere la durata del viaggio d'andata o di quello di ritorno?

An02: sì, sia l'uno / sia l'uno che l'altro

Woz: roma - milano: la durata del viaggio è tre ore e quarantacinque minuti.
milano - roma: la durata del viaggio è tre ore e trentacinque minuti

An02: va bene

Woz: se vuoi puoi procedere con l'acquisto del biglietto tramite carta di credito, altrimenti puoi prenotare il biglietto

An02: sì voglio acquistare il biglietto

Woz: non riesco a comprendere, puoi ripetere?

An02: voglio acquistare, i biglietti

Woz: va bene, per l'acquisto dovresti dirmi quale carta hai e il suo numero

An02: è carta visa. sette quattro nove sette, zero zero cinque cinque sei tre quattro due

Woz: va bene, aspetta un attimo. hai acquistato un biglietto per roma-milano il ventisette aprile alle dieci e quaranta, e un biglietto di ritorno per il ventotto aprile alle nove e trenta. ti serve altro?

An02: no grazie

Woz: ciao

An02: ciao

6.3 La gestione del turno nel *corpus* Speaky

6.3.1 Introduzione

La possibilità di usare la voce non basta a caratterizzare le interfacce vocali come human-like (cfr. 3.4.3) e a sfruttare la metafora della conversazione per l'interazione con un sistema (cfr. 3.4.1.2). I sistemi concepiti come human-like (e più facilmente identificabili come interlocutori), oltre a lasciare maggiore libertà sulla forma con cui presentare una richiesta, consentono un tipo di interazione in cui lo svolgimento della conversazione (e dell'attività) varia in base al contributo più o meno attivo dell'utente. Tale spazio di azione è definito dalle restrizioni previste per la successione dei turni (cfr. 2.4.1) e della gestione del passaggio dei turni.

La successione dei turni è determinata dalla gestione dell'iniziativa prevista e dal tipo di output usati: un sistema a iniziativa singola, ad esempio, obbliga il parlante a seguire un'architettura fissa di turni, spesso strutturata come una serie di richieste del sistema volte a elicitare informazioni per il completamento di un'attività (cfr. capitolo 4); inoltre, i sistemi a iniziativa singola prevedono grammatiche di comprensione restrittive (cfr. 2.3) per ogni stato del dialogo, a cui si associano richieste dirette (*direct prompt*) che riducono la gamma di risposte possibili alla sola informazione richiesta dal sistema: la sequenza dei turni è quindi altamente prevedibile, visto che il contributo del parlante è limitato al contenuto dei turni più che all'organizzazione della conversazione. Nell'interazione con un sistema a iniziativa mista l'utente non è rigidamente costretto dalla richiesta del sistema e può dare un maggiore contributo nell'indirizzare lo svolgimento dell'attività; per questo, sono usate anche richieste aperte (*open prompt*), che, oltre ad ampliare l'insieme di possibili risposte dell'utente, gli cedono l'iniziativa per lo svolgimento dell'attività: la sequenza di turni è quindi meno prevedibile, essendo legata anche alle strategie pragmatiche del parlante.

Anche la gestione del turno contribuisce a configurare la conversazione: a seconda di come viene preso e ceduto il turno di parola, possono variare le sequenze dei contributi e la composizione dei singoli turni: ad esempio, i sistemi che prevedono il *push-to-talk* affidano la gestione del turno ad azioni extralinguistiche, mentre quelli che prevedono il *barge-in* (ovvero, la possibilità di interrompere il

turno del sistema, cfr. 6.3.3.3) consentono le sovrapposizioni da parte dell'utente.⁴¹

Sia la gestione dell'iniziativa sia la gestione del turno concorrono ad avvicinare il sistema a uno dei due modelli (il sistema come interfaccia/tool-like e il sistema come interlocutore/human-like, cfr. 3.4), determinando la rilevanza di alcuni tratti della situazione comunicativa, come il formato di produzione conversazionale, che può essere deciso *a priori* dall'architettura del sistema, o la dimensione temporale dell'interazione, che oscilla tra sincronia e asincronia, a seconda della gestione del turno prevista.

Nel capitolo, verrà evidenziato lo stretto legame tra la gestione dell'iniziativa e la gestione del turno, che comporta diversi tipi di sequenze nelle fasi della conversazione meno strettamente legati alla risoluzione del compito e alle indicazioni del sistema. Verranno inoltre analizzati i passaggi del turno e illustrati i possibili errori legati a una gestione del turno affidata al parametro del silenzio, evidenziando sia la variabilità nei tempi di risposta al sistema a seconda del soggetto coinvolto e dell'esperienza acquisita, sia il legame tra la reattività dei soggetti e il tipo di output detto dal sistema.

6.3.2 Architettura dei turni e gestione dell'iniziativa

Come già scritto in 3.2.2, il tipo di dialogo che generalmente caratterizza l'interazione con un sistema artificiale è chiamato dialogo pratico (Allen et al. 2000), per via della stretta correlazione tra scopi dell'interazione e svolgimento di un'attività, per cui il sistema è programmato.

Il macro-modello della conversazione uomo-macchina può essere suddiviso in cinque fasi: 1) saluti 2) apertura 3) negoziazione 4) chiusura 5) saluti (Patzols et al. 1995, Alexandersson et al. 1997). I saluti iniziali coincidono con l'identificazione del parlante e l'attivazione del sistema; successivamente, nella fase di apertura viene individuata l'attività che il sistema e l'utente svolgeranno nella fase di negoziazione (in cui si confrontano il modello di svolgimento del sistema e le intenzioni del parlante) e che giunge a compimento nella fase di chiusura, in cui il sistema restituisce al parlante il feedback dell'attività svolta o l'informazione richiesta, a cui può seguire la chiusura del canale comunicativo, tramite i saluti di commiato, o l'apertura di un'altra attività. Se il macro-modello, essendo molto generico, è

⁴¹ Bisogna precisare che i sistemi *push-to-talk* spesso consentono il *barge-in*: tuttavia, l'interruzione non avviene tramite la sovrapposizione al turno del sistema, ma attraverso un'azione fisica.

piuttosto indipendente dal tipo di sistema coinvolto, le possibili sequenze di turni interne alle singole fasi sono condizionate dalla gestione dell'iniziativa prevista nel sistema di dialogo, dal tipo di architettura prevista per lo svolgimento dell'attività e dal tipo di output.

La sequenza di saluti, che serve a stabilire il contatto tra il parlante e il sistema, è iniziata dal parlante se il sistema necessita l'attivazione tramite un comando vocale apposito, altrimenti viene iniziata dal sistema se questo si attiva tramite azioni extralinguistiche (come una telefonata a un centralino o la pressione di un pulsante). Nella fase di apertura, l'attività da svolgere può essere individuata dal sistema (se, ad esempio, è prevista una sola attività, come nel sistema del *corpus* C-ORAL-ROM, cfr. cap. 4) o dal parlante, se il sistema è in grado di svolgere più attività e chiede al parlante di individuarne una (in genere con una domanda aperta, come nel *corpus* Speaky). La fase di negoziazione può essere rigidamente guidata dal sistema, che, tramite domande dirette e costrittive, completa il modello per lo svolgimento dell'attività; al contrario, se il sistema è a iniziativa mista, le domande possono essere aperte o chiuse (ma con grammatiche non ristrette al singolo stato del dialogo), portando a un diverso numero turni della fase di negoziazione a seconda delle informazioni fornite dal parlante.

6.3.2.1 Architettura dei turni del sistema nel *corpus* Speaky

Poiché il sistema è pensato come assistente casalingo attivabile principalmente via comandi vocali⁴², ogni interazione del *corpus* inizia con un input vocale del parlante, composto da un saluto o dal nome del sistema (spesso associati, 6.3.2.2), a cui segue il saluto del sistema e la richiesta di indicazioni sul tipo di attività da svolgere:

Ciao, in cosa posso esserti utile?

Ciao, come posso aiutarti?

Ciao, ti serve qualcosa?

Con la richiesta aperta successiva al saluto, il sistema cede l'iniziativa al

⁴² Per il sistema vero e proprio sarà possibile anche un'attivazione *push-to-talk* tramite il dispositivo acustico: la pressione del pulsante sostituisce il comando di attivazione vocale, per cui l'interazione inizia direttamente con il saluto del sistema.

parlante, che può così aprire la parte dell'interazione dedicata allo svolgimento dell'attività, rispondendo alla richiesta e indicando il dominio e l'attività da svolgere. La conversazione entra così nella fase di negoziazione, che si svolge su due livelli: il confronto linguistico, condotto attraverso la negoziazione tra ciò che il parlante dice e ciò che il sistema comprende (e viceversa), e il confronto pragmatico, che si svolge attraverso la negoziazione tra come il parlante intende svolgere l'attività e come il sistema è programmato per svolgerla. Nella fase di negoziazione possiamo quindi distinguere tra output di gestione dell'attività e output di gestione della comprensione.

I primi sono le richieste di completamento del sistema, volte a ricavare dal parlante le variabili necessarie per la risoluzione del compito:

AN18: <ee>, prenotazione nuove analisi

WOZ: dimmi quando e in quale ospedale vuoi prenotare le analisi

AN18: il venti dicembre, otto di mattina

WOZ: ciao, ti serve qualcosa?

AN16: sì <ee> volevo prendere un antibiotico che si chiama augmentin

WOZ: ok, quando devi prendere l'augmentin?

AN16: <ee> per una settimana, e vorrei che tu mi<vv> ricordassi di prenderlo

WOZ: vuoi essere avvisato con un sms, con un avviso vocale o con un
avviso sonoro?

NV04: un avviso vocale

Gli output di gestione della comprensione, che verranno approfonditi nel capitolo successivo (cfr. 6.4), sono le richieste di ripetizione e le richieste di conferma:

WOZ: ciao, in che cosa posso esserti utile?

AN23: <ee> volevo sapere le offerte che ci sono<vv> questa settimana alla
coop

WOZ: non ho capito, puoi ripetere?

AN23: sì, volevo sapere le offerte, che ci sono, al supermercato coop

NV05: devo andare a milano, andata ventisette aprile di mattina, ritorno
ventotto aprile di mattina

WOZ: vuoi un treno per merano?

NV05: per milano

L'iniziativa mista e la domanda aperta all'inizio di ogni attività consentono al parlante di organizzare lo svolgimento dell'attività senza essere costretto a rispondere unicamente alle richieste del sistema. Il peso informativo del turno di apertura dell'attività e dei successivi turni della fase di negoziazione determina una diverso numero di turni e diverse sequenze, sempre iniziate dal sistema:

WOZ: ciao, in che cosa posso esserti utile?

AN12: devo impostare la sveglia, dal lunedì al venerdì, alle sette del mattino

WOZ: vuoi che si accenda alle sette ogni giorno della settimana?

AN12: dal lunedì al venerdì

WOZ: va bene, la sveglia è stata impostata alle sette di mattina tutti i giorni della settimana ad esclusione del weekend. ti serve altro?

WOZ: ciao, in che cosa posso esserti utile?

AN16: allora, dovrei mettere<vv> la radio sveglia, alle sette di mattina

WOZ: puoi dirmi i giorni in cui vuoi che la sveglia si accenda all'orario indicato?

AN16: dal lunedì al venerdì

WOZ: vuoi che si accenda alle sette, ogni giorno della settimana?

AN16: dal lunedì, al venerdì

WOZ: va bene, la sveglia è stata impostata alle sette di mattina tutti i giorni della settimana ad esclusione del weekend

La lunghezza e la composizione della fase di negoziazione è determinata inoltre dalla presenza di incomprensioni e di errori. Nel caso di mancata acquisizione è prevista una sequenza di risoluzione di due turni (cfr. 6.4.5); nel caso di errate acquisizioni, la sequenza iniziata dalla richiesta di conferma del sistema può essere composta da due turni o da quattro turni, a seconda della strategia di correzione

adottata del parlante (cfr. 6.4.6):

WOZ: di quanti euro vuoi ricaricarlo?

AN22: di venti euro

WOZ: vuoi fare una ricarica di cento euro?

AN22: no di venti euro

WOZ: di quanti euro vuoi ricaricarlo?

NV10: venti

WOZ: vuoi fare una ricarica di cento euro?

NV10: no

WOZ: di quanti euro vuoi ricaricarlo?

NV10: venti euro

Le sequenze di turni della fase di negoziazione variano perciò a seconda delle variabili necessarie per lo svolgimento dell'attività, delle informazioni fornite del parlante e della capacità di comprensione del sistema.

Una volta fornite le informazioni previste per il completamento dell'attività, il sistema simulato restituisce nel turno finale, a seconda del compito, l'informazione richiesta (treni disponibili, film in programmazione, ecc.) o il feedback dell'attività svolta (spegnimento delle luci, impostazione della sveglia, ecc.): nella simulazione, tali feedback di completamento possono concludere il dialogo, nel caso l'attività sia l'ultima prevista per il compito, oppure non essere conclusivi, nel caso il compito preveda lo svolgimento di un'altra attività. In entrambi i casi, se il parlante non prende il turno per aprire l'attività successiva, dopo tre secondi dal feedback di completamento, il wizard chiede al soggetto se ha intenzione di svolgere un'altra attività:

WOZ: quando vuoi prendere il nuovo appuntamento?

AN21: il giorno sette dicembre, alle ore quindici

WOZ: va bene, ho preso un nuovo appuntamento per il sette dicembre alle quindici [2.8] posso fare qualcos'altro per te?

AN21: sì, vorrei<vv> ** che fossi avvisato

Nel caso l'attività sia l'ultima da svolgere all'interno del compito e il parlante rifiuti l'inizio di una nuova attività, il sistema saluta il soggetto, chiudendo il canale:

WOZ: va bene, aspetta un attimo. hai acquistato un biglietto per roma-milano il ventisette aprile alle dieci e quaranta, e un biglietto di ritorno per il ventotto aprile alle nove e trenta [3.1] ti serve altro?

AN12: no grazie

WOZ: ciao

Lo svolgimento e le sequenze di turni della fase di negoziazione sono perciò strettamente legate alle richieste del sistema e al contenuto informativo delle risposte del parlante, più che alla contrattazione del turno di parola.⁴³ La gestione del turno (e conseguentemente dell'iniziativa) è invece rilevante per la configurazione delle sequenze più indipendenti dalle richieste del sistema: la fase dei saluti di apertura/attivazione e di chiusura/disattivazione, e la fase di chiusura/apertura delle attività. In tali fasi, le diverse sequenze non sono determinate dal carico informativo del singolo turno ma dalle scelte conversazionali del parlante.

6.3.2.2 Saluti di apertura

La maggior parte dei parlanti seleziona una propria formula di attivazione, usata per tutta la sessione: *ciao*, *ciao speaky*, *speaky ciao* o il solo *speaky*.⁴⁴ Nella maggior parte dei casi, il primo output del sistema è identificato come un turno a sé stante, costituito dal saluto e dalla richiesta di apertura, a cui segue l'apertura dell'attività da parte del parlante, che segue perciò l'architettura prevista per l'attivazione:

AN10: ciao speaky

WOZ: ciao, in che cosa posso esserti utile?

AN10: <ee> impostare una radiosveglia

⁴³ Anche se, come verrà approfondito successivamente (cfr. 6.3.3.6.2), il parlante può provare a prendere il turno durante output che presentano una breve pausa all'interno, come nel caso delle richieste di ripetizione.

⁴⁴ Rilevando l'omogeneità delle formule di attivazione, bisogna tenere conto che ai soggetti è stato detto che il sistema si attivava chiamandolo per nome o salutandolo.

AN01: speaky

WOZ: ciao, in che cosa posso esserti utile?

AN01: come impostare la radiosveglia

NV03: ciao

WOZ: ciao, in che cosa posso esserti utile?

NV03: vorrei impostare la radiosveglia, per le ore sette di mattina, nei giorni
dal lunedì al venerdì

Un gruppo di parlanti (7 soggetti, tutti anziani) non identifica l'output iniziale del sistema come un singolo turno, di cui bisogna aspettare il completamento per prendere parola, bensì come una sequenza di due turni. Infatti, la breve pausa, che nell'output del sistema separa il saluto dalla richiesta di indicazioni sull'attività da svolgere, viene sfruttata per prendere il turno: l'output fisso, preregistrato, risulta così scomposto in due turni di conversazione. In questi casi, la fase di saluto/attivazione è composta da una tripletta: nel primo turno viene chiamato il sistema⁴⁵, a cui segue la coppia di saluti iniziata dal sistema e chiusa dal parlante:

AN23: speaky

WOZ: ciao

AN23: ciao+

WOZ: +in che cosa posso esserti utile?

AN23: eh senti vorrei delle informazioni sul mio conto in banca, il saldo
precisamente

AN08: speaky

WOZ: ciao

AN08: ciao # buongiorno #

WOZ: # in cosa posso # esserti utile?

AN08: <ee> sì, ho bisogno di mettere <ee> la radiosveglia, alle sette di
mattina. dal lunedì al venerdì

⁴⁵ In tutte le occorrenze, eccetto due casi, nel primo turno il sistema non viene salutato ma solo chiamato con il suo nome.

AN13: ciao speaky

WOZ: ciao

AN13: ciao

WOZ: in cosa posso esserti utile?

AN13: recupero vecchia prescrizione delle analisi che ho fatto a giugno

La tripletta, che deriva da una diversa identificazione del turno del sistema (non coincidente con l'output preregistrato) viene talvolta abbandonata nel corso della sessione: l'equivalenza tra turno e output del sistema può essere “stimolata” dall'esperienza con il sistema; infatti, la chiusura della coppia di saluti da parte dell'utente nella breve pausa interna all'output può causare sovrapposizioni (cfr. 6.3.3.6) di cui il sistema non tiene conto, a differenza del parlante che è costretto a interrompersi:

AN05: speaky

WOZ: ciao

AN05: # ciao #

WOZ: # in # che cosa posso esserti utile?

AN05: io vorrei recuperare la prescrizione delle analisi che ho fatto a giugno

AN05: speaky

WOZ: ciao,

AN05: ciao # s* #

WOZ: # ti # serve qualcosa? #

AN05: sì sono fuori fuo* / sono fuori e vorrei chiudere l'acqua di casa

In una delle sessioni, dopo quattro dialoghi iniziali in cui la fase dei saluti è composta dalla tripletta, il parlante abbandona il saluto e si limita ad attivare il sistema con la semplice allocuzione.

In un altro caso la sequenza iniziale a tripletta viene abbandonata subito dopo il primo compito: una volta compresa l'estensione del turno non negoziabile, il parlante si adatta e abbandona la tripletta. Dei cinque soggetti che non abbandonano la presa del turno durante l'output del sistema, tre soggetti alternano le due diverse sequenze (coppia o tripletta), lasciando come formula d'attivazione il solo nome del

sistema.

Tra le formule di attivazione va segnalato un *lapse* che può essere indicativo di quanto detto a proposito della rilevanza della copresenza spaziale nell'interazione con un sistema artificiale (cfr. 3.2), che, a prescindere dall'effettiva copresenza fisica, può diventare irrilevante per l'utente, essendo irrilevante per il sistema. È il caso di *pronto* (durante la sessione di AN_03), il segnale di apertura del canale comunicativo tipico delle conversazioni telefoniche, corretto immediatamente con una delle formule canoniche del *corpus*:

AN02: pronto <ee> ciao speaky

WOZ: ciao, ti serve qualcosa?

AN02: mi dovresti fare un favore, ce* cercare quali sono i film di woody allen, e acquistarli

6.3.2.3 Saluti di chiusura

Tre secondi (circa) dopo aver restituito il feedback di completamento dell'attività finale, il sistema chiede al parlante se vuole iniziare un'altra attività: in caso di rifiuto, il sistema saluta il parlante. A differenza dei saluti di apertura, più immediatamente associabili dal parlante alla funzione di apertura del canale comunicativo, la risposta al saluto di chiusura del sistema spesso non è vista come un comando funzionale: il suo saluto, infatti, può essere interpretato di per sé come un segnale di chiusura del canale.

Per questo, la fase dei saluti di chiusura è più eterogenea e variamente distribuita tra i soggetti, a differenza della sostanziale omogeneità della fase dei saluti di apertura. I tipi di sequenza possono essere suddivisi tra quelli iniziati dal parlante, che saluta subito dopo il feedback di completamento del sistema o successivamente alla richiesta di apertura di una nuova attività, e quelli iniziati dal sistema, a cui possono seguire i saluti del parlante o, in altri casi, il silenzio.

La principale sequenza di chiusura è composta dal saluto del sistema a cui segue il saluto del parlante (47%):

WOZ: ti serve altro?

NV09: no grazie

WOZ: ciao

NV09: ciao

WOZ: ti serve altro?

AN03: no grazie

WOZ: arrivederci

AN03: ciao

L'altra tipologia ampiamente rappresentata nel corpus (38%) è composta solo dai saluti del sistema, a cui non segue la risposta del parlante, probabilmente perché non è ritenuta funzionale per l'interazione. Particolarmente rappresentative sono due sessioni in cui i soggetti, dopo i primi dialoghi in cui rispondono al saluto del sistema, abbandonano successivamente la risposta:

WOZ: ti serve altro?

AN14: no grazie

WOZ: ciao

WOZ: ti serve altro?

NV12: no

WOZ: ciao

A evidenziare la stretta correlazione tra la scelta di salutare il sistema e la funzione che può essere attribuita al saluto, valgono alcuni commenti metalinguistici raccolti durante l'interazione (da parte di due soggetti non vedenti). Nel primo caso il parlante chiede informazioni per sapere se il *ciao* serva a chiudere il canale o se il canale si chiude automaticamente dopo il saluto del sistema:

WOZ: ti serve altro?

AN14: no grazie

WOZ: ciao

AN14: ciao. [bisogna rispondere al ciao o è finita lì?]

Nel secondo caso, il parlante aspetta il saluto del sistema per chiudere il canale:

WOZ: ti serve altro?

NV03: [aspetta aspetta deve di' ciao]

WOZ: ciao

NV03: [ecco, no perché così non riceve più comandi] ciao

Una sequenza meno frequente (9%)⁴⁶, ma comunque preferita da quattro parlanti, è composta dai saluti iniziati dal parlante successivamente al feedback di completamento del sistema o successivamente alla richiesta del sistema, a seguito del rifiuto:

WOZ: bene. il ticket è stato pagato

AN23: grazie. ciao

WOZ: ciao

WOZ: va bene, la radiosveglia si accenderà a volume massimo e rimarrà
accesa per quindici minuti

AN10: ciao

WOZ: ti serve altro?

AN10: no ciao grazie

WOZ: ciao

In tale sequenza, il saluto del parlante trova una funzione come feedback di corretta realizzazione dell'attività e come modo per velocizzare l'interazione. Uno dei soggetti, ad esempio, prova nei primi dialoghi della sessione ad “anticipare” l'ultima domanda del sistema (come nel primo esempio); successivamente, vista l'inefficacia di tale strategia, prova ad anticipare i saluti, facendoli seguire al feedback di completamento dell'ultimo sottocompito (come nel secondo esempio):

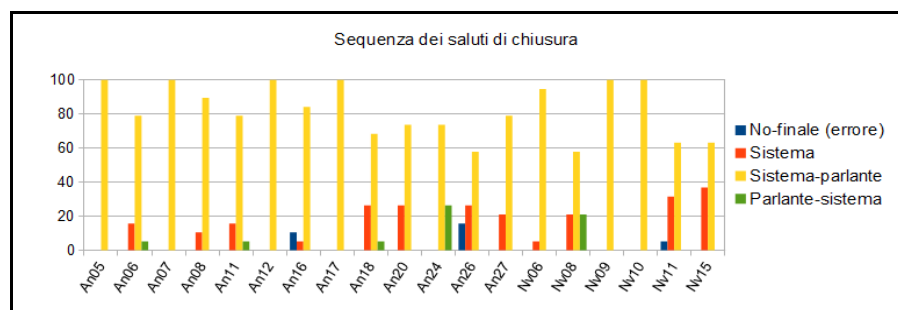
WOZ: va bene, ho prenotato al ristorante i tre moschettieri alle dieci per due
persone

⁴⁶ Le percentuali delle varie sequenze non danno come somma cento; è presente, infatti, un numero di casi (6%) in cui, per via di errori del wizard, non è stato seguito il protocollo di chiusura e, successivamente feedback di completamento, non è stata formulata la richiesta di apertura di una nuova attività.

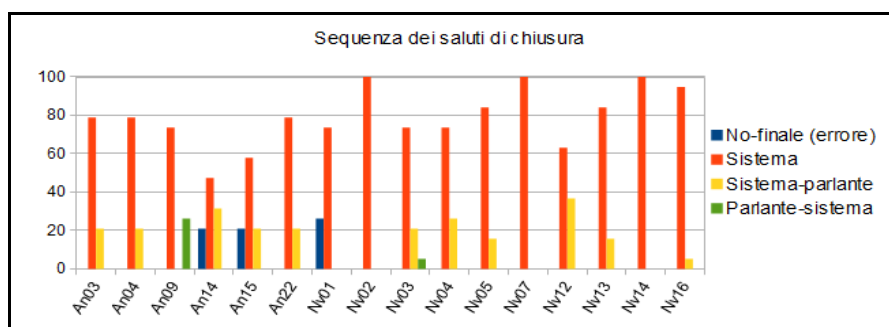
AN10: non mi # serve altro #
 WOZ: # ti serve altro? #
 AN10: no grazie
 WOZ: ciao

WOZ: va bene, hai ricaricato il tuo cellulare di venti euro
 AN10: non mi serve altro ciao
 WOZ: ciao

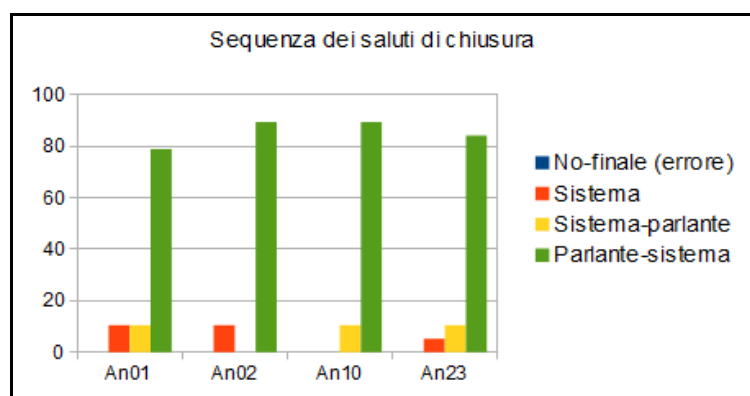
Complessivamente le sequenze dei saluti di chiusura sono variamente rappresentate nelle sessioni dei diversi parlanti. All'eterogeneità interpersonale delle sequenze di saluto, si contrappone una sostanziale omogeneità per ogni soggetto: come si può vedere nei grafici successivi (7, 8, 9), suddivisi per preferenza (saluti del sistema iniziali e risposta del soggetto, saluti del sistema senza risposta e saluti del parlante iniziali), una delle tipologie, esclusi i casi di errore (cfr. nota 46), è ampiamente maggioritaria per ogni parlante, rispetto ad altri tipi di sequenze (comunque presenti):



Graf. 7: Distribuzione delle sequenze di saluti tra i parlanti che salutano dopo il saluto del sistema



Graf. 8: Distribuzione delle sequenze di saluti tra i parlanti che non rispondono al saluto del sistema



Graf. 9: Distribuzione delle sequenze di saluti tra i parlanti che non rispondono al saluto del sistema

6.3.2.4 Feedback di completamento

Una volta terminata la fase di negoziazione, il sistema restituisce al parlante un output finale di svolgimento dell'attività, con l'informazione richiesta o il feedback del completamento di un'attività. Ogni compito è composto da almeno due attività: dopo la chiusura delle attività non conclusive del compito, il parlante può aprire la nuova attività di sua iniziativa o aspettare le indicazioni del sistema (cfr. 5.6.3.4). Il protocollo di interazione del wizard prevede infatti la possibilità per il parlante di avviare di sua iniziativa una nuova attività, da esercitare entro i tre secondi successivi al feedback di completamento, al termine dei quali è prevista la nuova richiesta da parte del sistema per sapere se il parlante vuole iniziare una nuova

attività.

Successivamente feedback di completamento del sistema possiamo trovare tre diverse sequenze di turni: apertura da parte del parlante, attesa del parlante e richiesta del sistema, feedback di accordo del parlante e richiesta del sistema. Dall'analisi delle transizioni tra un'attività e un'altra in tutto il *corpus* (complessivamente 1204), risulta una maggioranza di casi con iniziativa del parlante: la sequenza preferenziale è infatti composta dal feedback di completamento e dalla successiva presa del turno e apertura del parlante, presente nel 60% delle transizioni di attività:

WOZ: va bene, la temperatura della camera da letto è stata impostata a
ventitre gradi

AN10: adesso, se puoi mi puoi impostare l'accensione dello scaldabagno,
tutte le mattine si deve accendere alle, cinque e si deve spegnere, alle
nove

WOZ: va bene, lo scaldabagno si accenderà alle cinque di mattina e si
spegnerà alle nove. ti serve altro?

WOZ: ho prenotato le analisi per il venti dicembre allo otto del mattino
presso l'ospedale san camillo

AN08: perfetto, adesso me fai una cortesia, devi informarti sul costo delle
analisi

WOZ: costa settantacinque euro

L'attesa delle indicazioni del sistema porta invece a due configurazioni diverse di sequenza dei turni: il parlante può infatti restituire un feedback di accettazione dell'azione del sistema (che segnala inoltre il successo dell'interazione) e cedere il turno al sistema, oppure non restituire alcun feedback e non prendere il turno, rimanendo in attesa della una nuova richiesta. Nelle transizioni di attività l'iniziativa del sistema è meno frequente: la sequenza feedback di completamento-richiesta di apertura del sistema è presente nel 31% dei casi, mentre la sequenza feedback di completamento-feedback del parlante-richiesta di apertura del sistema è presente nel 9% delle transizioni:

NV03: voglio informazioni sul mio / sul saldo, in banca

WOZ: il saldo disponibile è di tremilacinquecento euro. ti serve altro?

NV03: sì, vorrei ricaricare il cellulare di venti euro

WOZ: hai preso il biglietto per lo spettacolo delle ventuno e trenta

AN28: va bene

WOZ: posso fare qualcos'altro per te?

AN28: <ee> vorrei prenotare anche il taxi

Suddividendo i feedback di completamento tra feedback di un'attività svolta (come l'accensione delle luci) e output contenenti un'informazione richiesta dal parlante (come i treni previsti per una determinata tratta) si può notare una diversa distribuzione di casi di iniziativa del parlante; infatti, nel caso di output di informazione, la successiva apertura del parlante (sequenza feedback di completamento-apertura del parlante) è presente nel 72% dei casi, mentre l'apertura del sistema (sequenza feedback di completamento-apertura del sistema) è presente nel 23% dei casi, seguita dalla sequenza feedback di completamento-feedback del parlante-apertura del sistema, presente nel 5% dei casi:

WOZ: ciao, come posso aiutarti?

NV14: <ee> mi / volevo sapere quali sono i film registrati, ieri

WOZ: vuoi registrare un film?

NV14: <vv>no volevo<vv> sapere quali sono i film registrati

WOZ: ieri ho registrato due film: casablanca e indovina chi viene a cena

NV14: <ee>. volevo scegliere casablanca, da vedere

WOZ: ciao, come posso aiutarti?

AN04: vorrei [cough] sapere l'elenco dei film registrati

WOZ: vuoi registrare un film?

AN04: no, vorrei l'elenco dei film registrati

WOZ: ieri ho registrato due film: casablanca e indovina chi viene a cena. ti serve altro?

AN04: sì, vorrei scegliere un film

WOZ: ciao, come posso aiutarti?

AN28: ho bisogno dell'elenco dei film registrati dalla tv
WOZ: vuoi registrare un film?
AN28: no voglio vedere l'elenco dei film registrati ieri
WOZ: ieri ho registrato due film: casablanca e indovina chi viene a cena
AN28: ok
WOZ: vuoi vedere un film registrato?
AN28: sì
WOZ: quale film vuoi vedere?
AN28: indovina chi viene a cena

Nei casi in cui la transizione è successiva all'output che segnala un'attività svolta, l'iniziativa del parlante è presente nel 49% dei casi, quella del sistema nel 41% dei casi, mentre il solo feedback di comprensione nel 10% dei casi:

WOZ: va bene ho prenotato le analisi per il venti dicembre alle otto del
mattino presso l'ospedale san camillo
AN13: senti puoi dirmi anche il costo?
WOZ: va bene, ho prenotato le analisi per il venti dicembre, alle otto del
mattino, presso l'ospedale san camillo. ti serve altro?
AN27: sì il pagamento del ticket
WOZ: ho prenotato le analisi per il venti dicembre, alle otto del mattino,
presso l'ospedale san camillo
NV13: grazie
WOZ: ti serve altro?
NV13: sì, pagare il ticket

La diversa distribuzione dell'iniziativa del parlante a seconda dell'output del sistema è legata al collocamento dei due tipi di attività all'interno dei compiti scritti per la simulazione: la richiesta di informazioni è sempre seguita da un'attività da svolgere in relazione all'informazione fornita (come negli esempi visti prima, in cui la richiesta dell'elenco è funzionale alla selezione di uno dei film); al contrario, il legame tra diverse attività da svolgere è meno stretto (come si può vedere negli

esempi precedenti).

Quando non è prevista alcuna attività successiva e il compito è concluso, successivamente feedback di completamento del sistema sono presenti solo due tipi di sequenza: l'attesa dell'output di apertura del sistema (a cui segue la disconferma del parlante) o un feedback di accordo (che può essere un vero e proprio feedback o un ringraziamento), a cui segue comunque la richiesta di apertura del sistema e la successiva disconferma. Nel *corpus* le due sequenze *grosso modo* si equivalgono; la prima è presente nel 49% dei casi e la seconda nel 51% dei casi:

WOZ: bene, il ticket è stato pagato. ti serve altro?

AN26: no grazie

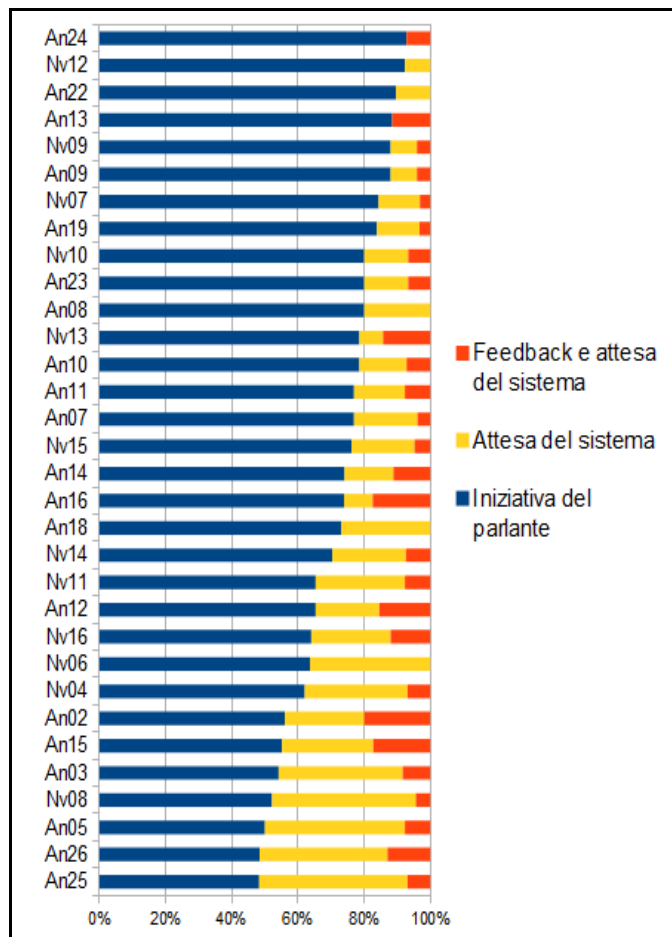
WOZ: va bene, la richiesta di fattura per entrambe le sedute è stata inviata

NV15: grazie

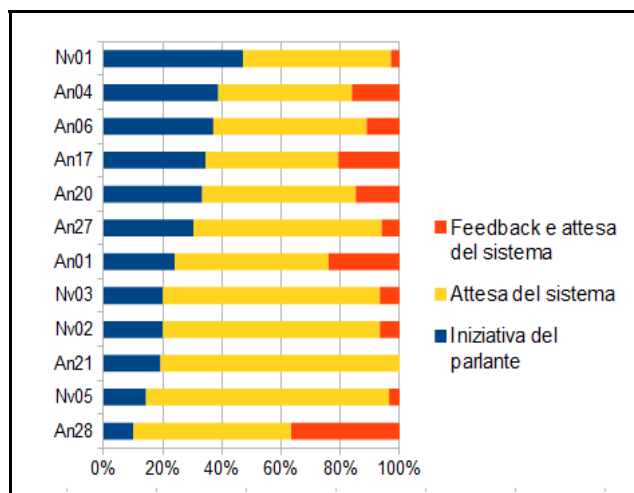
WOZ: ti serve altro?

NV15: no nono

La maggior parte dei parlanti preferisce aprire l'attività successivamente al turno di chiusura del sistema; tuttavia, il singolo soggetto durante la sessione può impiegare entrambe le strategie. Nel complesso, di 44 soggetti, 32 preferiscono prendere il turno e aprire l'attività successiva (graf. 10), mentre 12 preferiscono lasciare l'iniziativa al sistema (graf. 11).



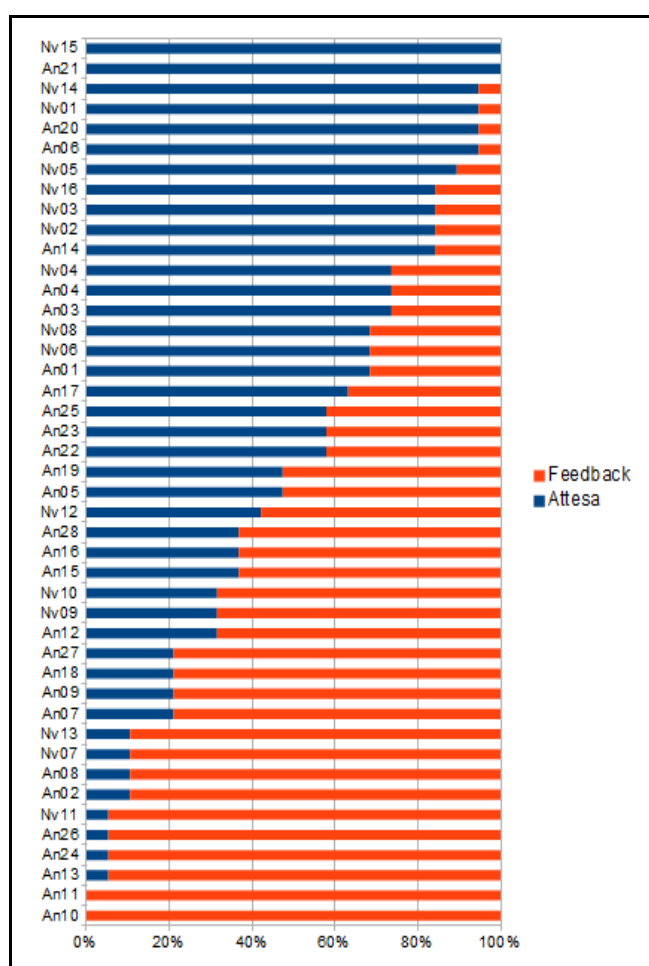
Graf. 10: Distribuzione delle sequenze successive al feedback di completamento tra i soggetti che preferiscono prendere l'iniziativa



Graf. 11: Distribuzione delle sequenze successive al feedback di completamento tra i soggetti che preferiscono non prendere l'iniziativa

Distinguendo tra i due tipi di completamento (feedback di un'attività svolta o informazione richiesta), la distribuzione è in linea con quanto detto precedentemente. Successivamente all'informazione fornita dal sistema, 35 parlanti preferiscono prendere l'iniziativa mentre 9 la lasciano al sistema; successivamente al feedback di un'attività svolta, 23 parlanti preferiscono prendere l'iniziativa mentre 21 preferiscono lasciarla al sistema.

Quando l'attività è l'ultima del compito, la preferenza tra le due scelte (feedback o attesa) è sostanzialmente equivalente: 20 parlanti preferiscono non dare alcun feedback al sistema, mentre 24 restituiscono un feedback al sistema; in entrambi i gruppi la preferenza può avere una gradazione diversa a seconda del parlante, come mostrato nel grafico (graf. 12).



Graf. 12: Distribuzione per parlante della sequenza iniziata dal feedback di completamento del sistema

6.3.3 La gestione del turno

Come per il dialogo uomo-uomo, la co-costruzione del dialogo nell'interazione uomo-macchina si basa sull'avvicendamento dei turni di diversi parlanti, tuttavia, l'alternanza dei turni viene in gran parte stabilita a priori e non viene gestita a livello locale e interazionalmente. Nella conversazione tra persone, infatti, la contrattazione del turno incide sul contesto e sulla *facies* dei singoli turni: da una parte ogni contributo alla conversazione è *context-shaped*, ovvero è determinato dal contesto, e allo stesso tempo è *context-renewing*, ovvero determina il contesto (Heritage 1985). Nell'interazione con un sistema di dialogo, non solo l'iniziativa lasciata dalla macchina determina quanto il contributo del parlante sia determinato dal contesto e quanto lo determini, ma anche la gestione del turno usata dal sistema, che può determinare sia le sequenze dei turni, sia la lunghezza e la composizione di questi (come vedremo in 6.3.3.6). Nei paragrafi precedenti abbiamo visto come alcune sequenze conversazionali siano strettamente correlate alla presa del turno da parte del parlante: le sequenze a tripletta dei saluti sono dovute a una presa del turno durante l'output del sistema; le differenti sequenze dopo il completamento di un'attività sono determinate dall'attesa del parlante o dalla sua presa del turno (da esercitare entro tre secondi). Il tempo di presa del turno da una parte è condizionato dalla rilevanza che il parlante dà alla sincronia (non automaticamente ricavabile dalla copresenza fisica: gli output possono essere visti come messaggi registrati), dall'altra, questa è fortemente condizionata da come il sistema gestisce la presa e la cessione del turno.

6.3.3.1 La gestione del turno nella conversazione uomo-uomo

Il turno non è definito dalla presenza di elementi ricorrenti che lo caratterizzano (come parole, frasi, enunciati, ecc.); infatti, è un'unità della conversazione che viene determinata interattivamente dagli interlocutori, tramite meccanismi di alternanza dei contributi. In altri termini, il turno è il tempo a disposizione dei parlanti per parlare, costruito con una o più unità linguistiche (per cui si parla di *Turn Constructional Unit*), mentre la sua composizione “linguistica” è contingente a ciò che dicono i parlanti nella conversazione. Sacks et alii 1974

descrivono il turno come un'unità che 1) richiede delle dimensioni minime, le quali tuttavia possono essere espanse all'interno dell'unità stessa; 2) può non essere completata; 3) possiede dei luoghi di passaggio ricorrenti al suo interno, che a loro volta possono essere contratti o espansi. La completezza (o quella che viene percepita come completezza) del turno di conversazione è definita dai TRP (*Transitional Relevant Place*, 'punto di rilevanza transizionale'), ovvero i punti in cui operano le regole di allocazione del turno successivo. Secondo le regole individuate da Sacks et alii (1974) il passaggio del turno può avvenire tramite due tecniche fondamentali: la prima prevede che il turno successivo sia allocato da chi detiene il turno attraverso la selezione del parlante successivo, la seconda prevede che il turno successivo sia preso dal parlante successivo che si autoseleziona. Appena la prima unità costitutiva del turno è completata o sta per essere completata, se la tecnica è quella del “parlante che seleziona il parlante successivo” allora è dovere del parlante selezionato prendere il turno in quel preciso luogo; se la tecnica è quella dell'autoselezione, nel punto di rilevanza transizionale un altro partecipante può prendere il turno; nel caso la conversazione coinvolga più interlocutori vale la regola per cui il primo che comincia ha il diritto di parola. Sempre relativamente alla tecnica dell'autoselezione, il parlante che detiene il turno può continuare a parlare, a meno che qualcun altro non si autoselezioni. Se il parlante non ha selezionato il parlante successivo e nessuno si è autoselezionato dopo il primo punto di rilevanza transizionale, il parlante che detiene il turno può continuare a parlare e il sistema di gestione del turno si applica nuovamente a partire dal successivo punto di rilevanza transizionale. La selezione del parlante successivo può essere esplicita, tramite l'uso di vocativi, nomi propri e direzione dello sguardo, altrimenti implicita, quando i requisiti per la presa del turno limitano tale azione a un solo partecipante (Lerner 2003).

La gestione locale di questo sistema di regole (ovvero un'allocazione che non è determinata a priori) rende inoltre conto dei fatti comuni della conversazione elencati da Sacks et alii 1974: un parlante alla volta, il numero non predefinito di partecipanti alla conversazione, la lunghezza non predeterminata del turno, i silenzi e le sovrapposizioni. Relativamente a questi ultimi due eventi conversazionali tipici, le regole permettono ai parlanti di interpretare il silenzio successivo a un punto di rilevanza transizionale; può infatti essere un “vuoto” (breve silenzio prima di un autoselezione, sia del parlante successivo sia del parlante precedente che continua il

turno), un “indugio” (nel caso in cui nessuno viene selezionato e nessuno si autoselezioni) o un silenzio con un suo preciso significato pragmatico (nel caso il parlante selezionato non prenda il turno). La gestione locale consente inoltre una gestione delle sovrapposizioni: se si verifica una sovrapposizione, uno dei parlanti generalmente si ritrae rapidamente; la persona che è stata designata per parlare ripete la parte del turno sovrapposta e quindi potenzialmente non comprensibile; se uno dei parlanti non si ritrae rapidamente, conquista il turno chi è “in ascesa”, una tecnica che coincide con l'innalzamento del tono, il rallentamento del ritmo, l'allungamento delle vocali.

6.3.3.2 La gestione del turno nelle interfacce grafiche

Nei sistemi a interazione vocale attuali l'alternanza dei turni di conversazione è basata su una gestione del turno molto robusta e per questo molto restrittiva, per certi versi somigliante all'interazione con le interfacce grafiche più che all'interazione tra persone: al comando del parlante (come può essere un clic del mouse) inviato tramite uno spazio d'interazione disponibile (come l'ambiente grafico) corrisponde la risposta del sistema relativamente a quel comando: il contributo dell'utente nell'interazione viene esercitato all'interno di un'architettura definita *a priori*. Oltre a definire lo spazio possibile d'interazione, l'interfaccia grafica definisce anche la tempistica dei contributi dell'utente: il sistema, tramite messaggi di attesa, icone (come la clessidra) o la barra di completamento di un'operazione (durante un'installazione), segnala all'utente quando non può essere eseguito alcun comando ed è necessario attendere. Invece di concorrere all'organizzazione della conversazione, la sovrapposizione è un errore che può verificarsi quando il sistema non restituisce il feedback di attesa o lo restituisce in ritardo: ad esempio, se all'azione su un oggetto non corrisponde un feedback, può accadere che l'utente esegua nuovamente l'azione, anche se il sistema sta già operando.⁴⁷ Le sovrapposizioni con le interfacce grafiche, inoltre, non sono intenzionali: la ripetizione del comando, da parte dell'utente, non è funzionale alla

⁴⁷ Un altro aspetto della gestione del turno che riguarda le interfacce grafiche è il completamento automatico (previsto per programmi di videoscrittura o form da compilare sul web). Il completamento è simile, come *facies*, al completamento del turno da parte dell'interlocutore frequente nelle interazioni umane; tuttavia, non ne condivide altre funzioni fondamentali: ad esempio, non coincide con un presa del turno da parte del completatore, né indica l'avvenuta acquisizione dell'intero enunciato, come avviene nella conversazione.

presa del turno, ma è causata dall'assenza di feedback, che viene intesa come problema del canale comunicativo.⁴⁸

6.3.3.3 La gestione del turno nelle interfacce vocali

Il passaggio del turno può essere più o meno vicino, da una parte, a quello dell'interazione con le interfacce grafiche, rigidamente definito dal sistema, dall'altra all'interazione tra persone, in cui il tempo di parola non è prefissato ma contrattato durante la conversazione. La distanza dalla gestione del turno delle conversazioni tra persone è definita dalle modalità previste del sistema, che possono anche concorrere contemporaneamente.

1) Modalità *push-to-talk*: il canale comunicativo tra persona e sistema si apre tramite la pressione di un apposito tasto e si chiude con il rilascio: in tale segmento di tempo può essere detto l'input. Con altri dispositivi, la pressione e il rilascio del tasto attivano il canale comunicativo, che viene chiuso una volta finito l'input e individuato un segmento di silenzio (come in Siri, cfr. 2.5). In entrambi i casi la cessione del turno del sistema e la presa del turno del parlante sono ottenuti con un'azione extralinguistica. Gli errori relativi a tale gestione del turno sono legati alla scarsa consuetudine dei parlanti a prendere il turno tramite un'azione extralinguistica: il turno linguistico può iniziare prima che il pulsante venga premuto e il canale aperto; oppure, l'input può finire dopo che è stato rilasciato il pulsante e il canale chiuso: anche l'assenza di una sillaba può portare problemi al modulo di comprensione.

2) Segnale acustico: il *push-to-talk* è usato principalmente nei sistemi in cui è il parlante a avviare la conversazione; quando l'interazione parte con una domanda del sistema, o nei turni successivi, in cui il sistema rivolge domande al parlante, il sistema può segnalare la cessione del turno con un segnale acustico che coincide con l'attivazione del canale e la possibilità di dare un input. Anche in questo caso, i problemi sono legati alla scarsa consuetudine con tale gestione del turno: i parlanti spesso iniziano a parlare ancora prima del segnale acustico, essendo abituati a prendere il turno una volta finito il turno dell'interlocutore (Boyce 2008).

⁴⁸ Probabilmente, una maggiore esperienza con un'interfaccia comporta una minore presenza di questo tipo di sovrapposizioni: si ha infatti una maggiore conoscenza delle situazioni che possono verificarsi, e, nello specifico, una maggiore consapevolezza che un'assenza di feedback non è correlata automaticamente a un problema del canale comunicativo.

3) Modalità VAD (Voice Activity Detection): alcuni sistemi consentono una gestione del turno più somigliante a quella usata quotidianamente; la presa e la cessione del turno non avvengono tramite azioni extralinguistiche: la sequenza procede secondo la regola del “un parlante alla volta” e durante il turno del sistema il canale è disattivato e il sistema non “ascolta”. Quando finisce il turno del sistema, se il parlante comincia a parlare si attiva il canale comunicativo, che viene chiuso quando il sistema prende nuovamente il turno⁴⁹. La presa del turno da parte del sistema si basa generalmente sull'individuazione di un segmento di silenzio successivo al completamento del turno che segnala che il turno è finito; questo può portare a sbagliare l'interpretazione di alcune pause. Anche in questo caso, la differenza dalla gestione umana può causare, anche se in misura minore, alcuni errori, ad esempio se il parlante inizia il suo turno sovrapponendosi alla conclusione del turno del sistema.

4) *Barge-in*. Attualmente, la modalità più “umana” per la gestione del turno è il *barge-in*: durante il turno del sistema il canale comunicativo resta aperto, per cui il parlante può interrompere il sistema e prendere il turno per inviare l'input successivo. Il *barge-in* consente un'interazione più veloce e meno legata ai tempi imposti dal sistema, tuttavia meno robusta: il canale rimasto aperto durante il turno del sistema può far aumentare il numero di errori, dovuti a rumori accidentali (come un colpo di tosse o una macchina) o a feedback di comprensione del parlante, interpretati come presa del turno da parte del parlante invece che segnali di corretta acquisizione (Beringer et al. 2001).

6.3.3.4 Voice Activity Detection

La modalità usata per la gestione del turno del sistema Speaky è la Voice Activity Detection (cfr. 5.6.2), che consente di verificare se l'utente sta parlando. La rilevazione continua è basata sull'energia del segnale audio: una volta stabilito un livello d'intensità base del segnale (che tenga conto del rumore di fondo), il canale si attiva una volta individuato un segnale di forza maggiore del livello base e si

⁴⁹ La cessione del turno basata sulla conclusione del turno del parlante, individuata tramite un breve periodo di silenzio, è usata talvolta anche nei sistemi che prevedono il push-to-talk solo per prendere il turno.

disattiva una volta che il segnale torna sul livello precedente. Il rilevamento viene condotto su microfinestre temporali (10ms), il cui segnale audio viene successivamente etichettato come parlato rivolto alla macchina o come assenza di input da parte del parlante. Per ridurre i problemi relativi al rumore di fondo, il rilevamento è basato su un valore chiamato SNR (*signal-to-noise ratio*), ovvero il rapporto tra segnale dell'input e rumore di fondo: una volta che il valore supera un livello preimpostato, si attiva il canale comunicativo. Naturalmente, la collocazione del livello base può determinare la presenza maggiore o minore di falsi positivi e falsi negativi: un valore alto porterà a una maggiore presenza di mancate ricezioni da parte del sistema, un valore più basso a una maggiore presenza di input inesistenti.

Nel caso in cui il sistema consenta il *barge-in*, la VAD rimane operativa durante il turno del sistema, a cui il parlante si può sovrapporre per interromperlo e consentirgli così di saltare parte del turno del sistema. L'assenza del *barge-in* irrobustisce l'interazione (togliendo la possibilità di errori durante il turno del sistema) ma può essere frustrante, soprattutto nel caso di interazioni o output lunghi e ripetitivi. La presenza del *barge-in* porta con sé gli stessi errori causati dalla VAD: il sistema può ad esempio interrompersi a fronte di un rumore, di un feedback del parlante, di un colpo di tosse, processando quindi il segnale acquisito come comando. Una possibile soluzione è impostare un'attesa prima dell'interruzione del turno del sistema, per estrarre nuove caratteristiche dal segnale del parlante (o quello detto o se ancora deve essere completato) e validare l'interruzione, nel caso abbia senso in sé e in relazione allo stato del dialogo; d'altra parte, se il sistema mantiene il turno a fronte di un reale *barge-in*, il parlante può pensare di non aver preso il turno e interrompere il comando, che di fatto risulterà incompleto, o ripetere nuovamente il comando.

6.3.3.5 Rilevamento della cessione del turno

Il sistema deve anche saper rilevare la cessione del turno da parte del parlante, che nell'interazione uomo-macchina coincide con il completamento del comando relativo allo stato del dialogo. Per individuare il completamento, la VAD, oltre a rilevare una diminuzione del rapporto tra segnale e rumore, si basa anche sulla lunghezza temporale del segmento di silenzio successivo al comando: un silenzio superiore al limite imposto viene considerato come una cessione del turno. Anche in

questo caso, la collocazione del livello base di durata può portare a due tipi di errore (Raux 2008):

- *cut-in*: quando una pausa dell'enunciato del parlante è uguale o maggiore alla lunghezza imposta come parametro di cessione del turno, il sistema interpreta tale pausa come cessione del turno; inoltre, se il sistema non restituisce un feedback (linguistico o extralinguistico) dell'avvenuta ricezione e della chiusura del canale, durante l'ulteriore lasso di tempo impiegato dal sistema per processare il segnale, il parlante continua a parlare senza sapere di aver inviato un comando diverso da quello che il sistema sta processando. Ai problemi legati alla distanza tra l'input che il parlante è convinto di aver inviato e l'input in realtà processato dal sistema, si aggiunge il peggioramento dei dati del riconoscitore e del modulo di comprensione, che devono processare un comando parziale e quindi potenzialmente incompleto;

- *silenzi*: al contrario, quando il valore del segmento di silenzio impostato come parametro per la presa del turno da parte del sistema è troppo alto, l'eccessiva latenza tra l'input del parlante e la risposta del sistema, oltre a rallentare l'interazione, può essere interpretata dal parlante come una mancata ricezione; in tal caso, il parlante può tentare di ristabilire il canale comunicativo con la ripetizione dell'input o con segnali fatici di verifica del canale, che verranno interpretati dal sistema come parte del nuovo input.

Il silenzio come parametro per rilevare la cessione del turno da parte del parlante è legato alla difficoltà di definizione di completezza del turno. L'originaria caratterizzazione di Sacks et al. 1974, pur non esplicitamente, poneva la sintassi al centro dell'unità di turno, sulla base del passaggio al parlante successivo, spesso coincidente con la fine dell'unità sintattica («Examination of where such 'next-turn starts' occur in current turns shows them to occur at 'possible completion points.' These turn out to be 'possible completion points' of sentences, clauses, phrases, and one-word construction», p. 721); successivamente anche all'intonazione, in subordinazione alla sintassi, è stato riconosciuto un ruolo importante per il completamento del turno. Più recentemente è stato evidenziato come il passaggio del turno (e quindi la sua “completezza” in una prospettiva conversazionale) sia spesso non inquadrabile sulla base della completezza sintattica e prosodica. Nel passaggio e

nel mantenimento intervengono infatti molti aspetti legati al comportamento non verbale degli interlocutori coinvolti in uno specifico contesto (Ford – Thompson 1996) e in una specifico tipo di attività (Selting 2000), oltre che al comportamento verbale: hanno pari peso, infatti, connettivi pragmatici (ad esempio quelli di scansione del discorso, come *innanzitutto*, *poi*, *infine*, ecc.), gestualità, sguardo, ecc.⁵⁰ Ford – Thompson 1996 hanno analizzato due delle caratteristiche di completamento del turno individuate in Duncan 1972: completamento sintattico/grammaticale e intonazione finale (ascendente o discendente) e hanno rilevato che al completamento prosodico corrisponde quasi sempre il completamento sintattico (mentre non è così il contrario). Come sottolineano Gravano – Hirschberg 2011, il problema di tali analisi è che si confrontano solo con una parte dei completamenti, ovvero quando avviene effettivamente un passaggio di turno, mentre non considerano tutti gli altri spazi opzionali di passaggio del turno.

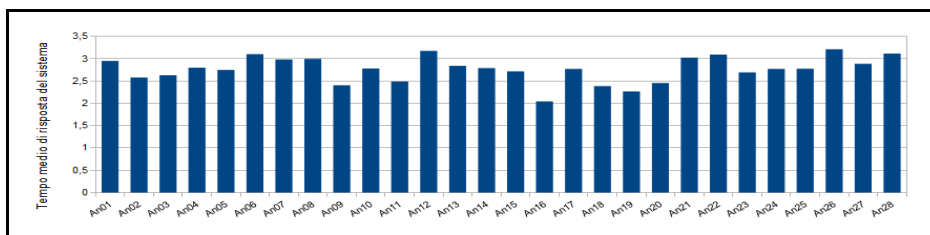
6.3.3.6 La gestione del turno nel *corpus* Speaky

6.3.3.6.1 I tempi di risposta del sistema

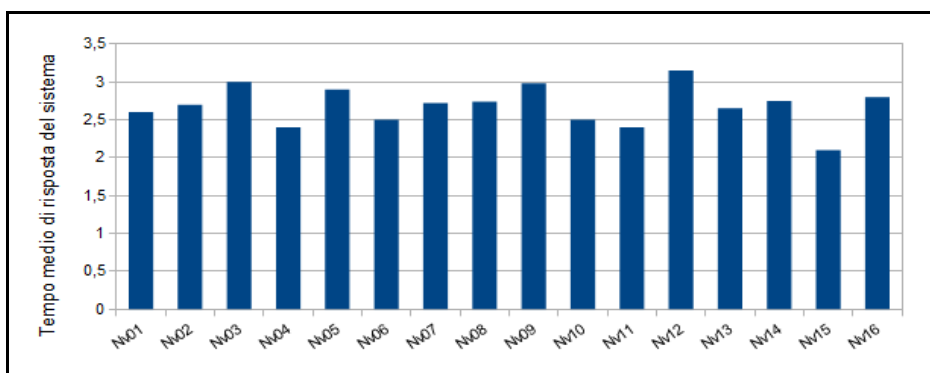
Il sistema Speaky prevede che il passaggio del turno dal parlante al sistema avvenga tramite l'individuazione di un segmento di silenzio alla fine dell'enunciato: nella raccolta del *corpus* è stata simulato un equivalente meccanismo di presa del turno, per avere una simulazione realistica della gestione del turno e per via di limitazioni tecniche dovute al tempo impiegato per scegliere un output. Per via delle numerose variabili difficilmente gestibili non è stato definito un tempo di risposta preciso nel protocollo per il wizard ma una finestra indicativa tra i due e i tre secondi.⁵¹ Nel *corpus* il tempo medio di risposta del sistema al parlante è di 2,7 secondi; questo può naturalmente variare a seconda dell'interazione e a seconda dell'input al sistema (vedi nota 49), tuttavia non ci sono stati particolari scarti tra le diverse sessioni, che oscillano tutte tra i due e i tre secondi di media:

⁵⁰ La stessa scomposizione in turni intesi come unità sintattiche/prosodiche può portare a non considerare molti tratti del comportamento verbale e non verbale e perciò non aiutare nella descrizione della conversazione.

⁵¹ Il tempo di risposta del sistema può variare a seconda di alcune variabili: la reazione del wizard all'input e la latenza del software una volta selezionato l'output, latenza da tenere maggiormente in considerazione quando si opera in rete (come abbiamo fatto per la raccolta a Torino). D'altronde, anche il tempo di risposta del sistema può variare a seconda del tempo impiegato per processare il segnale, stimabile ma non prevedibile con esattezza.



Graf. 13: Tempo medio di risposta medio del sistema simulato per ogni sessione (anziani)



Graf. 14: Tempo medio di risposta medio del sistema simulato per ogni sessione (non vedenti)

La gestione del turno ha portato a casi di sovrapposizione del sistema sul turno del parlante, affini al *cut-in*, ovvero a quei casi in cui il sistema interpreta come cessione del turno una pausa del parlante. Questo tipo di sovrapposizione, dovuto a un'errata presa del turno da parte del sistema, è tuttavia difficilmente distinguibile dall'altro tipo di sovrapposizione errata, dovuta a un'eccessiva latenza del sistema, che può spingere il parlante a riprendere il turno. Nel *corpus* sono presenti complessivamente 47 occorrenze di *cut-in* dovute a pause del parlante interpretate come cessioni del turno; il tempo medio di pausa del parlante nei casi in cui si verifica il *cut-in* è uguale al tempo medio di risposta del sistema, ovvero 2,7 secondi.

Il confine che intercorre tra i *cut-in* dovuti alla presa del turno del sistema (che non interpreta correttamente una pausa del parlante) e quelli dovuti a un'eccessiva latenza del sistema (che inducono il parlante a continuare il turno) è difficilmente formalizzabile da un punto di vista temporale. Tuttavia, bisogna rilevare che in nessun caso l'input viene ripetuto (come avviene generalmente quando il sistema ci mette troppo a rispondere, Ward et al. 2005), ma viene esteso, anche dopo

una lunga pausa. La latenza del sistema più che a un problema del canale comunicativo viene quindi ricondotta dal parlante all'incompletezza del comando (che infatti viene completato):

WOZ: il medico ti ha prescritto l'assunzione del fosamax70, per tre mesi una volta a settimana

AN24: allora

WOZ: il fosamax70 serve a ridurre i rischi dell'osteoporosi, come fratture vertebrali e dell'anca

AN24: ok. inseriscilo nella mia lista dei farmaci [2.5] il lunedì, alle # se* #

WOZ: # vuoi #
inserirlo nell'agenda dei medicinali da assumere?

AN01: mh vorrei sapere il costo, della<vv> del ticket

WOZ: costa settantacinque euro

AN01: perfetto pago con la carta visa [2.0] # quarantanove #

WOZ: # va bene #, per l'acquisto
dovresti dirmi quale carta hai e il suo numero

AN01: visa, numero quarantanove settantasette cinquanta cinquantasei diciotto ventidue

AN03: vorrei <ee> impostazione di un promemoria per l'assunzione di un farmaco

WOZ: va bene, puoi indicarmi il nome del farmaco?

AN03: augmentin

WOZ: ok, quando devi prendere l'augmentin?

AN03: tutta la settimana [2.1] dalle ore undici alle # quindici #

WOZ: # va bene #, puoi
indicarmi anche l'orario?

AN03: dalle ore undici alle quindici

Gli unici casi in cui la latenza del sistema è attribuita dal parlante a un problema del canale comunicativo sono presenti nella fase del saluto di apertura, che, come abbiamo visto, è la fase dedicata all'attivazione del canale comunicativo. La

latenza di risposta del sistema, non prevista nella simulazione e dovuta a errori (del wizard o del *software*), ricorre in 15 casi, e il tempo medio che intercorre tra il primo saluto e la ripetizione di questo è di 5,8 secondi.

Oltre ai possibili errori dovuti alla sovrapposizione del sistema sul turno del parlante per un'errata interpretazione di una pausa, nel *corpus* sono presenti casi (58) in cui il sistema si sovrappone al parlante perché quest'ultimo tenta di prendere il turno (ritraendosi immediatamente) durante la breve pausa presente in alcuni output, in genere comprendenti un feedback (di comprensione e svolgimento dell'attività o di incomprensione) e una richiesta. Alcuni parlanti, anche più volte nella stessa sessione, a fronte dello stesso tipo di output (come le richieste di ripetizione o i saluti già visti prima, cfr. 6.3.2.2), provano a prendere il turno prima della richiesta del sistema, interpretando la breve pausa del sistema come una cessione del turno:

WOZ: va bene, puoi dirmi la data e l'ora?

NV10: dunque, la data deve essere il quattordici dicembre alle undici di mattina la prima

WOZ: appuntamento prenotato per il 14 dicembre alle 11

NV10: [0.4] # la seconda #

WOZ: # vuoi fissare # un altro appuntamento?

NV10: sì le<vv> il ventuno dicembre alle undici

WOZ: appuntamento prenotato per il 21 dicembre alle 11

NV10: [0.3] # eh richiedi per cortesia #

WOZ: # vuoi fissare un altro # appuntamento?

NV10: no eh devi richiedere fattura per e* entrambe le sedute

AN20: prenoto da zia maria

WOZ: va bene, per quante persone?

AN20: otto persone

WOZ: non ho capito

AN20: [0.1] #o*#

WOZ: # puoi # ripetere?

AN20: otto persone

WOZ: va bene. hai acquistato un biglietto per roma-milano il 27 aprile alle 10 e 40, e un biglietto di ritorno per il 28 aprile alle 9 e 30

AN13: il mio nome lo vuoi?

WOZ: non ho capito

AN13: [0.1] il mio # no* #

WOZ: # puoi # ripetere?

AN13: il mio nome e cognome lo vuoi?

Un altro problema legato alla presa del turno che può verificarsi durante l'interazione è la partenza simultanea dopo un feedback di completamento, a cui segue dopo tre secondi circa la richiesta del sistema per sapere se deve svolgere un'altra operazione (cfr. 6.3.2.4): anche se tale situazione non è molto frequente (44 casi), si può ritrovare nella maggior parte delle sessioni (29 parlanti).

Al di là delle sovrapposizioni non intenzionali, durante il turno del sistema è molto raro che il parlante si sovrapponga o tenti di prendere il turno prima che l'output del sistema sia finito. Sono presenti solo 18 casi, tutti durante il completamento di un turno:

WOZ: ti serve # altro #?

NV14: # sì #, vorrei delle informazioni per regolare il riscaldamento visto che nei prossimi giorni ci saranno temporali e si abbasserà la temperatura

WOZ: ciao ti serve qualcosa?

NV15: sì. mi vai sulla guida dei film che <ee> faranno questa sera in televisione, perché ho deciso di non uscire

WOZ: vuoi vedere la programmazione sportiva per # stasera in tv? #

NV15: # no # la i film in programmazione dopo le ore venti

Sono rari anche i feedback di comprensione durante l'output del sistema, presenti solo in 13 casi, spesso durante turni lunghi composti da un elenco:

WOZ: quali vuoi selezionare?

AN17: l'olio. e il caffè lavazza

WOZ: hai selezionato: olio extravergine mancini da un litro, # a # 3 euro e 49 centesimi

AN17: # sì #

WOZ: # lavazza # decaffeinato intenso 250 milligrammi, # a # un euro e 89 centesimi

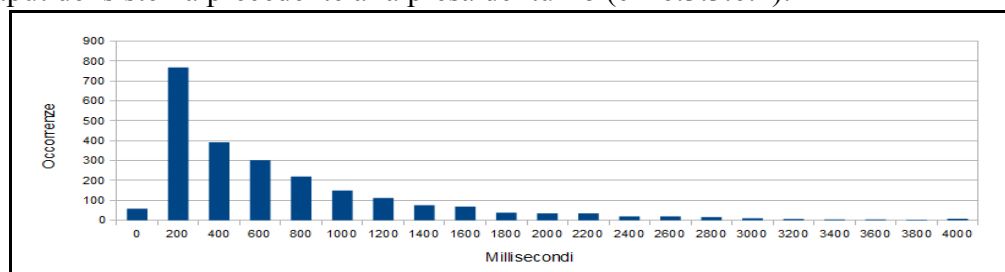
AN17: # sì #

sì. allora potresti ordinarmi

6.3.3.6.2 La presa del turno del parlante

Eccettuate le sovrapposizioni dovute alla poca esperienza con la gestione del turno del sistema, diversa rispetto a quella della comunicazione tra persone, e dei rarissimi casi di sovrapposizione e feedback intenzionali, le interazioni tra parlante e sistema sono principalmente caratterizzate dal silenzio sia durante il turno di conversazione del sistema sia durante il passaggio del turno.⁵²

Nel *corpus* il tempo medio di risposta dei parlanti agli output del sistema è di 0,7 secondi. Dei passaggi di turno analizzati, il 31% è inferiore agli 0,2 secondi (graf. 15) ; tuttavia tali tempi devono essere correlati, oltre che ai singoli parlanti, al tipo di output del sistema precedente alla presa del turno (cfr 6.3.3.6.4):



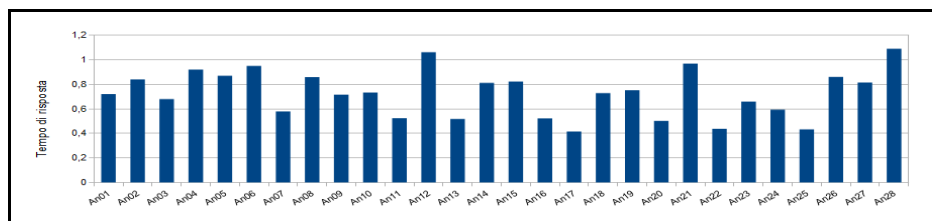
Graf. 15: Tempi di presa del turno da parte del parlante

⁵² Si veda a tal proposito l'importante studio di Heldner – Edlund 2010, che, oltre a passare in rassegna molte analisi dei tempi della conversazione, evidenzia, analizzando tre *corpus* in lingue diverse (Spoken Dutch Corpus, HCRC Map Task Corpus e Swedish Map Task Corpus) la grande frequenza di sovrapposizioni durante i passaggi del turno (sopra il 40%).

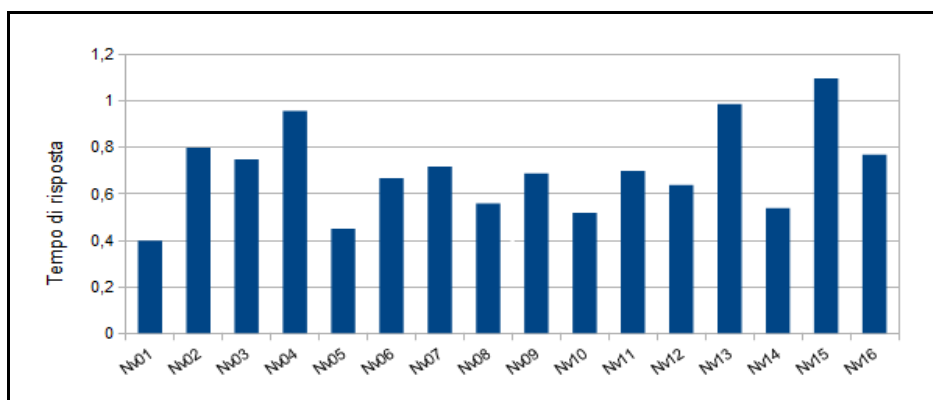
Complessivamente, contando anche i casi di sovrapposizione durante il passaggio del turno e l'assenza di silenzio tra i due turni, in tutto il *corpus* il 96,2% dei passaggi di turno è caratterizzato dal silenzio, mentre solo l'1,4% da sovrapposizioni e il 2,4% dall'assenza di silenzio o sovrapposizioni.

6.3.3.6.3 *Variazione interpersonale e intrapersonale*

Come già abbiamo visto, la gestione del turno può essere messa in relazione all'interpretazione data dal parlante alla dimensione temporale della conversazione: un tempo di risposta breve avvicina l'interazione alla sincronia delle conversazioni quotidiane; al contrario, un tempo di presa del turno più lungo denota una maggiore preoccupazione per facilitare la comprensione da parte del sistema (come avviene, ad esempio, per l'iperarticolazione dopo gli errori del sistema, cfr. 6.4.5.3.1). Il tempo di presa del turno varia così da soggetto a soggetto, con scarti piuttosto evidenti: come si può vedere nei grafici 16 e 17, il parlante più “lento” (An28) ha un tempo medio di risposta al sistema che è più del doppio del parlante più “veloce” (Nv01):



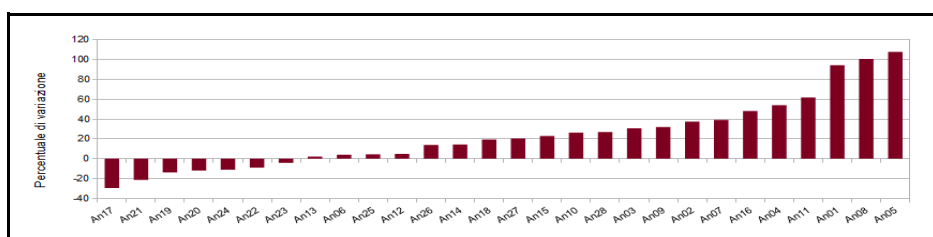
Graf. 16: Tempo medio di risposta medio dei soggetti anziani



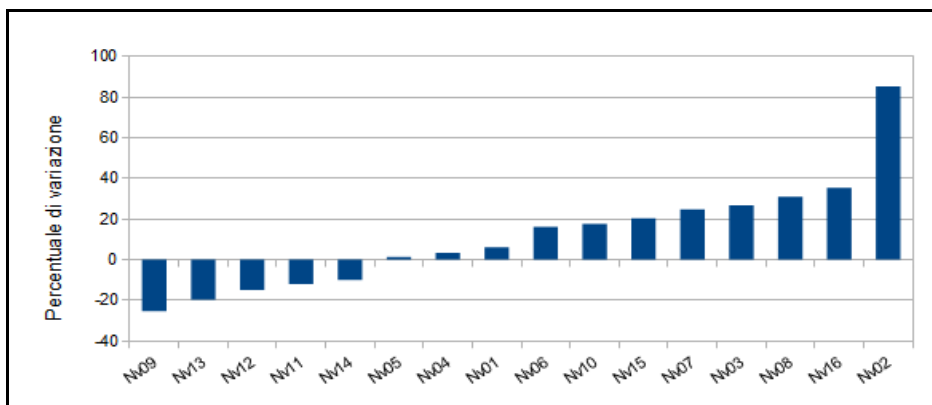
Graf. 17: Tempo medio di risposta medio dei soggetti non vedenti

Inoltre, visto il documentato adattamento dei parlanti ai tempi di presa del turno del sistema (Gustafson et al. 2008), e più in generale, la variazione che può comportare l'acquisita esperienza nell'interazione con il sistema di dialogo, sono stati messi a confronto i tempi di risposta dei primi tre compiti e quelli degli ultimi tre compiti: in tutto il *corpus* la parte finale della sessione ha dei tempi di risposta più lunga del 16%.

Come mostrato nei grafici (18 e 19), anche distribuendo i risultati tra i parlanti, la variazione tra i due momenti della sessione è poco consistente; tuttavia, per alcuni parlanti i tempi di risposta si allungano decisamente nella parte finale della sessione (arrivando ad aumentare di più del 100%), mentre è più raro che i tempi che si accorcino:



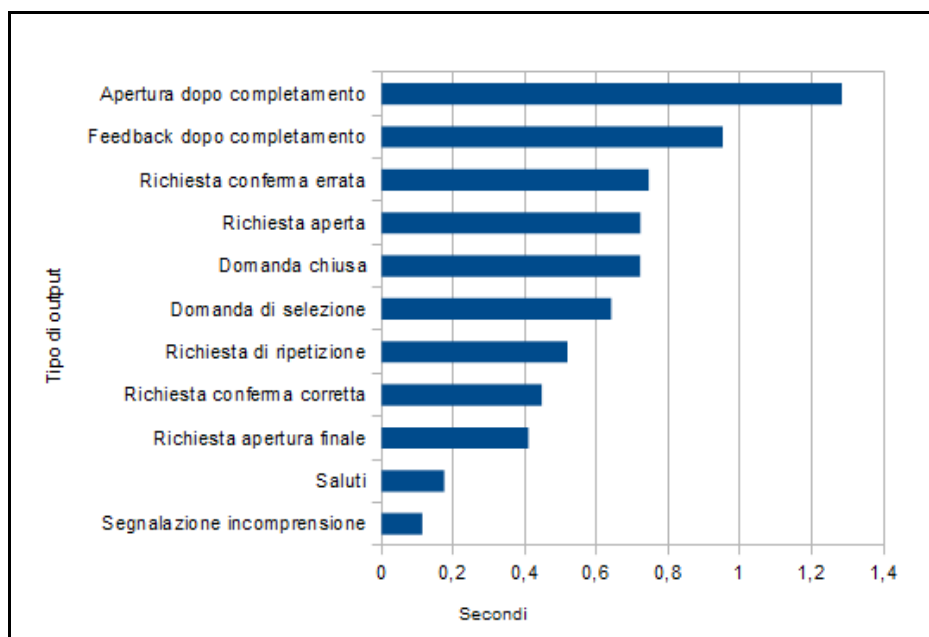
Graf. 18: Variazione del tempo medio di risposta dei soggetti anziani tra prima e seconda parte della sessione



Graf. 19: Variazione del tempo medio di risposta dei soggetti non vedenti tra prima e seconda parte della sessione

6.3.3.6.4 Tempi di presa del turno in relazione alla sequenza dei turni

Anche il tipo di sequenza dei turni può incidere sul tempo di presa del turno da parte del parlante. Isolando i tempi di risposta del parlante a seconda della tipologia di output del sistema precedenti alla presa del turno possiamo rilevare significative differenze a seconda dei turni coinvolti.



Graf. 20: Tempo medio di risposta dei soggetti in relazione al turno precedente del sistema

I due output dopo cui il parlante prende il turno più velocemente sono la segnalazione della mancata acquisizione (“non ho capito”, “non ho compreso”, che

compone solo una parte dell'output di ripetizione) e i saluti. Entrambi gli output, come abbiamo visto, sono talvolta interpretati diversamente dai soggetti rispetto alla reale lunghezza; in 78 casi il parlante prende il turno (nel caso dei saluti) o prova a prendere il turno (nel caso delle richieste di ripetizione) durante la breve pausa che separa la richiesta di ripetizione dalla segnalazione della mancata acquisizione, e il saluto dalla richiesta di apertura dell'attività.

Le sequenze di turni iniziate dal sistema presentano una pausa tra i due turni compresa tra i 0,4 e gli 0,7 secondi. Il tempo di presa del turno è più breve dopo l'apertura finale del compito (a cui non segue alcuna attività), che si differenzia per più di tre decimi di secondo dalla presa del turno successiva alla richiesta di apertura iniziale e intermedia del compito. La stessa differenza significativa si rileva tra le richieste di conferma corrette e le richieste di conferma sbagliate; mentre sono sostanzialmente equivalenti le domande dirette e le domande in cui il parlante deve selezionare un oggetto da un elenco.

Il tempo di presa del turno si allunga in maniera considerevole successivamente a un feedback del sistema, quando è assente una richiesta esplicita e il parlante può prendere l'iniziativa per aprire l'attività, sia nel caso venga dato un feedback di comprensione sia che venga aperto il nuovo sottocompito (cfr. 6.3.2.4)

6.3.4 Sommario e conclusioni

Nel capitolo abbiamo descritto la configurazione dell'interazione con il sistema simulato Speaky. Pur essendo determinate principalmente dall'output del sistema e dal carico informativo della risposta del parlante, le sequenze di turni, all'infuori della fase di negoziazione, sono legate esclusivamente alla presa del turno. Nel caso dei saluti di chiusura, la scelta è determinata dalla funzione, o dall'irrelevanza, che viene assegnata dal parlante al saluto per la chiusura dell'interazione; dopo feedback di completamento è invece legata alla gestione dell'iniziativa e al maggiore o minore controllo del parlante sullo svolgimento dell'interazione e dell'attività.

L'analisi dei tempi di risposta e dei tentativi di presa del turno mostra come l'interazione uomo-macchina sia caratterizzata principalmente dal silenzio, durante il turno del sistema e durante il passaggio del turno; inoltre, nella maggior parte dei casi, il turno di parola concesso al sistema coincide con la lunghezza dell'output,

senza la presenza di tentativi di presa del turno da parte del parlante. Tuttavia, durante alcuni output composti da più enunciati separati da una breve pausa (come ad esempio, il saluto iniziale o la richiesta di ripetizione) aumentano i tentativi di presa del turno, che possono comportare sovrapposizioni problematiche per il futuro sistema, nel caso verrà consentito il *barge-in*. Un altro genere di sovrapposizioni possono verificarsi nel caso di latenze del sistema, durante le quali il parlante invece di ripetere l'input, prova a completarlo: la distanza tra ciò che il parlante pensa di aver detto e ciò che effettivamente il sistema ha computato rischia di essere problematica per lo svolgimento dell'interazione.

Da un punto di vista applicativo, sono rilevanti i differenti comportamenti relativamente ai tempi di presa del turno, che possono variare da parlante a parlante e per lo stesso parlante (a seconda dell'esperienza acquisita), e possono segnalare differenti gestioni dell'iniziativa, nelle fasi con minore controllo da parte del sistema (come ad esempio, nell'apertura di un sotto-compito). Oltre a questa variabilità, per l'impostazione dei parametri relativi alla gestione del turno (tempi di attivazione del canale, scrittura degli output e delle sequenze di turni del sistema), bisogna tenere in considerazione il tipo di output del sistema, che può portare a diversi tempi di reattività da parte del parlante.

6.4 La gestione degli errori nel *corpus* Speaky

6.4.1 Introduzione

Come già visto nel capitolo 1, nell'interazione con un sistema di dialogo si presentano frequentemente situazioni d'errore; per questo, sin dai primi sistemi a iniziativa singola, è stata prestata particolare attenzione alla progettazione di algoritmi per la gestione degli errori. Un sistema, anche se molto semplice e con un'architettura costrittiva, può sbagliare l'interpretazione di un input per varie ragioni: la principale, e tuttora la più frequente, è l'errata decodifica del segnale acustico da parte del riconoscitore vocale. Con l'aumentare dell'elasticità dell'architettura e della libertà lasciata al parlante di formulare la propria richiesta, sono aumentate anche le cause che viziano il processo di comprensione, cause che possono afferire ad altri livelli linguistici oltre a quello acustico, come ad esempio quello semantico e pragmatico. Inoltre, l'evoluzione dei sistemi e la loro maggiore "naturalzza" indirizzano gli stessi utenti a essere meno formali e meccanici nella produzione dei propri enunciati, portando a una maggiore presenza dei tradizionali fenomeni legati ai tempi del parlato (come esitazioni, false partenze, parafrasi infraturno, pause piene, nonché la più vasta categoria dei segnali discorsivi), aumentando di fatto i potenziali problemi per il sistema.⁵³

Con i sistemi a iniziativa mista (2.4.1) la ricerca deve confrontarsi con nuovi problemi da risolvere nell'ambito della gestione degli errori, sia per dotare le grammatiche di comprensione di ulteriori caratteristiche da estrarre durante l'interazione che facilitino l'individuazione e la risoluzione di un errore, sia per avere informazioni su ciò che il parlante pensa della comprensione della macchina. Infatti, da una prospettiva incentrata sull'utente, l'analisi delle situazioni di errore, oltre a fornire un supporto conoscitivo per integrare le grammatiche di comprensione durante tali situazioni, fornisce informazioni utili sul parlante e sulla sua interpretazione del contesto. Se nelle conversazioni quotidiane è normale che gli interlocutori collaborino nella risoluzione di un problema di comprensione, controllando continuamente la reciproca comprensione⁵⁴; nel dialogo uomo-

⁵³ Per un approfondimento su ciò che comporta una maggiore naturalzza e "umanità" del sistema in relazione alla presenza di tali fenomeni rimandiamo nel parlato dell'utente rimandiamo al capitolo 3, oltre a Bell et al. 2000 e Bickmore – Cassell 2005 relativamente ai singoli fenomeni.

⁵⁴ Oltre alla reciproca collaborazione, spesso negoziata a seconda della situazione comunicativa di

macchina, pur essendo ridotta la casistica delle strategie di correzione (ogni sistema è infatti dotato di un modulo di gestione degli errori che “incanala” il parlante su procedure maggiormente controllate), l’interazione con un oggetto non umano e di cui non si conosce il funzionamento spinge il parlante a dover valutare come funziona il processo di comprensione della macchina per risolvere l’errore (Fischer 2003), comportandosi di conseguenza nella riformulazione del proprio turno:

«The sequential positions of the utterances produced as solutions to the same problem reveal the speakers' hypotheses about what causes the communicative problems because they constitute reformulations of the previous utterances. They thus provide not only interpretative cues as displays of the speakers' own interpretation of their previous utterances, but also display the speakers' concepts of the determining factors of the current situation. These involve at least the speakers' concepts about language, the robot as a communication partner, and the subject matter.» (Fischer 2003)

Durante la risoluzione dell’errore, oltre alle modificazioni dell’enunciato, l’interpretazione della situazione è resa ancor più evidente dai commenti metalinguistici (tra parentesi quadre, negli esempi) degli stessi parlanti, con cui questi esplicitano il ragionamento a proposito delle capacità di comprensione della macchina e di ciò che può aver causato la mancata acquisizione:

WOZ: vuoi pagare ora con carta di credito o pagare alla consegna?

AN03: pagarlo alla consegna

WOZ: non ho capito, # puoi ripetere? #

AN03: # pagarlo # pagarlo alla consegna [forse parlo troppo veloce]

WOZ: ciao, in che cosa posso esserti utile?

NV01: non ricordo quando ho l'appuntamento per la fisioterapia

WOZ: non ho capito, puoi ripetere?

cui si conoscono i possibili problemi (ad esempio, una conversazione telefonica disturbata) bisogna aggiungere che ogni parlante conosce *grossa modo* le capacità di comprensione di altri parlanti. Nei casi in cui queste potrebbero essere sconosciute, vengono spesso stereotipate (pensiamo a una conversazione con un parlante L2). La conoscenza delle capacità di comprensione di un sistema automatico è molto minore e si fonda su una scarsa esperienza.

NV01: <ee> a* non ricordo quando ho l'appuntamento con la fisioterapia
[forse ho capito]

WOZ: domani devi andare dal fisioterapista alle sedici

Purtroppo i commenti metalinguistici sono occasionali e non forniscono un quadro sufficientemente esauriente delle interpretazioni dell'errore. Nel presente capitolo verranno analizzati sia qualitativamente sia quantitativamente (laddove possibile) i vari tipi di variazione; si cercherà, inoltre, di associare i tipi di variazione all'interpretazione della situazione da parte dell'utente, allo scopo di evidenziare ciò che per il parlante può essere problematico, al di là del reale errore che può occorrere.

6.4.2 Tipologie d'errore nell'interazione con un sistema automatico

Durante il processo di comprensione il sistema decodifica ciò che è stato detto dal parlante e ne estrae le informazioni rilevanti in relazione al turno, al dominio dell'interazione e alle azioni che è in grado di eseguire. Il processo di decodifica non porta, perciò, a una vera e propria comprensione dell'enunciato, ma ad alcune ipotesi di input probabili, a seconda del segnale audio, della grammatica di comprensione prevista per il turno in questione e del compito da svolgere. Gli errori possono perciò occorrere in diversi passaggi del processo di acquisizione dell'input. Normalmente, il punto di partenza per la tipologizzazione dei potenziali errori è il modello di *grounding* proposto da Clark (Clark & Schaefer 1989, Clark 1996)⁵⁵. Per Clark i partecipanti a una conversazione stabiliscono un terreno comune (*common ground*) attraverso la collaborazione su quattro diversi livelli: 1) il parlante stabilisce un contatto con l'ascoltatore che a sua volta mostra di prestare attenzione; 2) il parlante produce un segnale che l'ascoltatore decodifica; 3) il parlante tramite il segnale dice qualcosa che viene compreso dall'ascoltatore; 4) il parlante tramite il segnale compie o richiede un'azione che viene accettata o integrata dall'ascoltatore. Il modello può essere facilmente sovrapposto ai passaggi del processo di ricezione e comprensione di un sistema, a cui possono essere collegate ad altrettante tipologie di

⁵⁵ Il modello è stato rielaborato da Paek 2003 nella sua tassonomia interdisciplinare di classificazione degli errori comunicativi; e successivamente ripreso da Bohus – Rudnicky 2005. Parallelamente sono state proposte simili tassonomie, sempre basate sull'approccio di Clark, in Larsson 2003 e Schlangen 2004.

possibili cause d'errore (prendendo come punto di riferimento la schematizzazione di Paek 2003).

6.4.2.1 Errori legati alla presa del turno e alla segmentazione dell'input. Nel dialogo uomo-macchina il turno, invece di essere continuamente negoziato, ha un valore discreto: è infatti equivalente a un input che viene dato al sistema dal parlante, simile a un clic col mouse o a una stringa di comando per un'interfaccia testuale. La gestione dei turni si basa solitamente sull'individuazione di una pausa (di lunghezza variabile a seconda del sistema), parametro che indica al sistema che il turno è completato e l'*input* può essere considerato completo, e quindi la ricezione essere chiusa⁵⁶. Gli errori, a livello conversazionale, possono verificarsi nel caso l'enunciato del parlante abbia lunghe pause e sia frammentato, portando il sistema ad acquisire un comando incompleto; allo stesso modo, la presenza di rumori di fondo, identificati dal sistema come parte del segnale audio, può portare il sistema a non prendere il turno (e di conseguenza, il parlante a ripetere o riformulare la propria richiesta), o a identificare il rumore come un enunciato, anche quando, ad esempio, il parlante è in silenzio.

6.4.2.2 Errori del riconoscitore vocale. Il passaggio dal segnale audio alla stringa trascritta è la fonte principale di errori (cfr. *infra*): l'articolazione delle parole può variare da persona a persona (e da situazione a situazione) e quindi non essere perfettamente adatta al modello acustico e statistico del sistema⁵⁷; il canale audio attraverso cui viene registrato il segnale è spesso viziato da rumori che possono alterare i valori estratti. Gli errori del riconoscitore vocale vengono usualmente suddivisi in *cancellazioni*, *inserzioni* e *sostituzioni*: le prime sono parti dell'enunciato che non vengono trascritte dal riconoscitore, le seconde sono parole presenti nella trascrizione ma non presenti nell'enunciato del parlante, le ultime corrispondono ai casi in cui viene decodificata una parola per un'altra. Per misurare l'affidabilità di un riconoscitore vocale si usa generalmente, come parametro, il Word Error Rate (WER), calcolato dividendo la somma dei tre tipi di errore per il totale delle parole da riconoscere. Tuttavia, non tutte le parti di un enunciato hanno la stessa rilevanza

⁵⁶ Nel caso sia previsto il *barge-in*, il parlante può prendere nuovamente il turno appena lo desidera; tuttavia, questo non integrerà automaticamente il precedente, ma, a seconda della risposta del sistema, lo cancellerà o ne integrerà la parte richiesta successivamente dal sistema.

⁵⁷ In questo caso, gli errori si presentano principalmente con sistemi che hanno riconoscitori vocali *speaker independent*, i cui modelli non si calibrano sull'utente specifico.

nell'interazione con un sistema di dialogo: ad esempio, nei sistemi con grammatiche *frame-and-slot* tutte le informazioni non rilevanti per il sistema non vengono inserite nelle grammatiche ma in una classe semantica chiamata *garbage*.⁵⁸ Per questo, talvolta, per valutare gli errori di un sistema vengono usati il CER (*Concept Error Rate*) o il KER (*Keyword Error Rate*), metriche che tengono conto solo delle parti rilevanti dell'enunciato in relazione al dominio e al compito che deve svolgere il sistema.

La maggiore o minore presenza di alcuni tipi di errore è inoltre determinata dalle stesse strategie di gestione degli errori basate sulla *confidence* (vedi *infra*): una gestione degli errori con una soglia di *confidence* alta può portare a un maggior numero di cancellazioni; al contrario, una soglia di *confidence* bassa e poco robusta può portare a una maggiore presenza di inserzioni.

6.4.2.3 Errori del modulo di comprensione. Un'altra tipologia di errori è dovuta a parole usate dall'utente ma non presenti nelle classi semantiche del sistema (OOV, *Out-of-vocabulary*) o a formulazioni dell'enunciato che non concordano con le grammatiche di comprensione scritte (OOG, *Out-of-grammar*). I primi si verificano quando il parlante usa una parola chiave che non è presente nelle classi semantiche del sistema, o se omette nella propria formulazione una parola chiave necessaria per l'esecuzione di un compito; i secondi sono dovuti a una diversa pianificazione del compito e delle azioni che si possono svolgere con il sistema in un determinato turno: tali errori possono verificarsi, ad esempio, se il parlante prende l'iniziativa quando non è previsto dalle grammatiche del sistema. Infine, l'errore può verificarsi per via di enunciati ambigui dal punto di vista semantico e pragmatico, che possono portare a un conflitto tra classi semantiche.

6.4.2.4 Errori del dominio d'applicazione. L'ultima tipologia di errori comuni con un sistema, più che da attribuirsi a un'errata comprensione della macchina, riguarda il dominio che gestisce il sistema e i compiti che può svolgere. Tali errori avvengono, infatti, quando il parlante chiede al sistema un'azione non coperta dal dominio del sistema (OOD, *Out-of-domain*) o nel caso in cui chieda di svolgere un'attività che il sistema non è in grado di svolgere (OOA, *Out-of-application*).

⁵⁸ La preferenza per le parole rilevanti rispetto a parole "vuote" è individuata dagli stessi utenti, che talvolta parlano al sistema eliminando ogni possibile fonte di disturbo e fornendo solo parole chiave (cfr. Danieli 2005).

6.4.3 Strategie di gestione degli errori

Per valutare l'acquisizione o il rifiuto di un input il sistema si affida a un parametro probabilistico: il riconoscitore vocale di un sistema di dialogo assegna a ogni parola decodificata un punteggio, chiamato *confidence*, che rappresenta la probabilità della parola trascritta di essere quella effettivamente pronunciata dal parlante. Ogni punteggio assegnato si colloca su una scala che va da 0 a 1, dove 0 rappresenta l'incomprensione completa, mentre 1 rappresenta la certezza assoluta della parola trascritta. Nella progettazione e nella valutazione di un sistema vengono posti dei “paletti” che definiscono la corrispondenza tra i valori di *confidence* e il comportamento del gestore di dialogo: più il valore è vicino a 0, più è opportuno che l'enunciato vada rifiutato e richiesto nuovamente; al contrario, più il risultato si avvicina a 1, più è probabile che l'enunciato sia stato individuato correttamente e vada quindi accettato senza chiederne conferma. La corrispondenza tra i valori di *confidence* assegnati dal riconoscitore e l'accettazione (o il rifiuto) di un determinato enunciato o parte dell'enunciato viene chiamata “soglia di *confidence*” (o “soglia di confidenza”). La soglia è variabile a seconda del sistema e viene definita in base alle sue caratteristiche. Ad esempio, un sistema con un vocabolario maggiore rispetto a un altro sistema ha una soglia di accettazione più alta, poiché la probabilità di decodificare correttamente un determinato enunciato è più bassa. Non solo il vocabolario determina la collocazione della soglia di *confidence*: un altro fattore è il tipo d'iniziativa prevista, che coincide con la libertà concessa al parlante relativamente a ogni stato del dialogo.

La collocazione della soglia sulla scala di *confidence* determina, inoltre, la robustezza di un sistema e la sua elasticità e naturalezza. Due sistemi di dialogo con gli stessi moduli e la stessa architettura, ma che differiscono per la collocazione della soglia di *confidence*, portano a due tipi di interazione diversa: il sistema con un limite più basso consente un'interazione più veloce e meno costrittiva per l'utente, d'altra parte il rischio che alcuni errori non vengano individuati subito è maggiore, il che comporta (successivamente) una maggiore difficoltà nella loro correzione e la possibilità che il compito non sia svolto correttamente e l'utente abbandoni l'interazione con il sistema. Dal lato opposto, il sistema di dialogo con una soglia di *confidence* più alta comporta un maggiore controllo su ogni informazione acquisita e

la minore probabilità che un errore non venga individuato subito dal sistema, al contempo una tale gestione degli errori allunga la conversazione per via dei numerosi turni di controllo e di conferma, portando a un'interazione che può essere frustrante per il parlante umano.

Le strategie di gestione degli errori non si limitano unicamente all'accettazione o al rifiuto di un enunciato: infatti, a seconda del grado di *confidence*, sono state elaborate diverse modalità di accettazione, di richiesta di conferma e di rifiuto. Le strategie, che verranno ora esposte in ordine decrescente a partire da un alto valore di *confidence*, non sono necessariamente presenti in ogni sistema; né, se presenti, devono essere attive durante ogni turno dell'interazione⁵⁹. Anche in questo caso, il modello che può essere preso come punto di riferimento è quello elaborato da Clark relativamente ai feedback di *grounding*: rilevanza del turno successivo, feedback di comprensione, parafrasi e ripetizione.⁶⁰ Al di sotto della soglia di *confidence* che definisce l'accettazione dell'enunciato esistono varie strategie per esplicitare ciò che il sistema ha capito, allo scopo di individuare possibili errori a seconda della reazione dell'utente.

6.4.3.1 Rilevanza e appropriatezza del turno successivo. Su livelli di *confidence* molto alti e prossimi alla certezza, il sistema può segnalare la corretta acquisizione dell'input unicamente facendo una richiesta successiva coerente con quanto detto dal parlante:

WOZ: vuoi prendere un altro appuntamento?

NV02: sì

WOZ: in che giorno?

NV02: sette dicembre

WOZ: puoi indicarmi l'ora?

NV02: quindici in punto

⁵⁹ Così come i primi sistemi di dialogo a iniziativa singola, gli attuali sistemi, in momenti particolarmente critici dell'interazione (come un'operazione bancaria, l'inserimento di dati personali, ecc.) possono prevedere un controllo fisso su ciò che è stato compreso dal sistema, a prescindere dal valore di *confidence*.

⁶⁰ Il modello di Clark prevede, innanzitutto, l'attenzione continuata, ovvero la dimostrazione che il canale è attivo e l'interlocutore sta ascoltando: per tale funzione, quando prevista nei sistemi di dialogo, è in genere previsto un feedback sonoro più che vocale, a cui corrisponde la cessione del turno e l'apertura della ricezione (così è previsto per Siri, per Google Voice e per il Voice Control del Windows Phone).

WOZ: ho preso un nuovo appuntamento per il sette dicembre alle quindici.
posso fare qualcos'altro per te?

Una strategia del genere può essere usata per dei livelli di *confidence* prossimi a 1, o quando un compito sia semplice e la successione di domande breve; infatti, essendo probabile, in tal caso, che un errore rimanga nascosto per molti turni, deve essere prevista, anche nei turni successivi, la possibilità di correggere l'acquisizione errata. Inoltre, per quanto la strategia comporti un'interazione agile e veloce, è possibile che il turno rilevante non basti a un utente diffidente come segnale di comprensione, portandolo a percepirlo come un possibile errore e spingendolo a richiedere l'esplicitazione di ciò che il sistema ha acquisito:

WOZ: va bene, per l'acquisto dovresti dirmi quale carta hai e il suo numero

AN17: è una visa. il numero è 74 97 00 55 63 42

WOZ: va bene, aspetta un attimo. hai acquistato un biglietto per roma-milano
il 27 aprile alle 10 e 40, e un biglietto di ritorno per il 28 aprile alle 9 e
30. ti serve altro?

AN17: volevo sapere se <ee> se pas* è andato a buon fine l'acquisto

WOZ: non ho capito, puoi ripetere?

AN17: volevo sapere se il numero della visa era giusto

Ci sono altre strategie che prevedono la non esplicitazione della comprensione, come ad esempio l'uso di segnali di accordo, come “va bene” e “ok”, prima degli output del sistema. L'uso di feedback di accordo, usuali nelle interazioni quotidiane, può in parte ovviare alla diffidenza dell'utente; non risolve, tuttavia, il problema dell'evidenza esplicita di ciò che è stato compreso dal sistema.

6.4.3.2 Richiesta di conferma implicita. Una modalità di conferma più robusta, che consente di non arrestare l'interazione sul turno problematico ma permette all'utente di proseguire nello svolgimento dell'attività, è la richiesta di conferma implicita, la quale è composta da un feedback esplicito di comprensione (che riprende l'informazione fornita dall'utente) e dalla richiesta successiva, con cui vengono chieste al parlante le informazioni necessarie per l'attività:

WOZ: va bene, puoi indicarmi il nome del farmaco?

NV12: augmentin

WOZ: ok, quando devi prendere l'augmentin?

NV12: tutta la settimana

Nel caso la risposta alla domanda del sistema sia coerente con quanto descritto, e non vengano identificati altre caratteristiche generalmente ascritte alle situazioni d'errore (cfr. *infra* e Krahmer et al. 2001), l'acquisizione viene considerata corretta e l'interazione procede: per questo, le grammatiche devono includere, nel turno successivo alla richiesta di conferma implicita, ulteriori metriche di valutazione dell'acquisizione. Gli attuali moduli di comprensione possono gestire una tale strategia solo su domini molto piccoli e con compiti molto semplici: una maggiore ampiezza di dominio e complessità delle azioni, a cui corrisponde una maggiore estensione delle grammatiche di comprensione, comporta il rischio che il sistema non riesca a individuare la correzione e la scambi per acquisizione, portando così a nuovi errori, che si aggiungono al precedente.

6.4.3.3 Richiesta di conferma esplicita. Per avere una gestione delle acquisizioni ancora più robusta possono essere usate le richieste di conferma esplicite, con cui il parlante viene obbligato a confermare o a disconfermare l'informazione compresa dal sistema. Da una parte l'uso delle richieste esplicite porta ad allungare l'interazione, dall'altra riduce di molto la possibilità che un errore non venga individuato dal sistema, costringendo il parlante a focalizzare l'attenzione su ciò che il sistema ha acquisito e di cui chiede conferma. Le grammatiche di comprensione devono essere programmate in funzione della semplice disconferma, a cui segue una nuova richiesta:

WOZ: va bene, di quanti euro vuoi ricaricarlo?

AN09: venti euro

WOZ: vuoi fare una ricarica di cento euro?

AN09: no

WOZ: di quanti euro vuoi ricaricarlo?

AN09: venti euro

Devono inoltre essere programmate in funzione di una disconferma a cui segue l'informazione corretta, consentendo così l'iniziativa mista (anche solo relativamente al turno di controllo):

WOZ: il credito disponibile è di tre euro e venti centesimi

AN22: vorrei fare una ricarica di venti euro, sul, mio cellulare

WOZ: vuoi fare una ricarica di cento euro?

AN22: no, di venti euro

6.4.3.4 Richiesta di ripetizione. La richiesta di ripetizione viene usata per segnalare al parlante che il sistema non ha acquisito alcuna informazione. A differenza delle strategie viste precedentemente, che possono essere presenti o meno nell'architettura del sistema, la richiesta di ripetizione è l'unica strategia di gestione dell'errore inserita in ogni architettura; infatti, alcuni tipi di errore (come errori di gestione del turno o richieste al di fuori delle attività consentite dal sistema) rendono possibile per ogni sistema la mancata acquisizione. D'altra parte, l'uso può essere esteso a valori di *confidence* più alti, preferendo la ripetizione (o riformulazione) completa rispetto alle correzioni parziali previste per strategie potenzialmente più efficaci e veloci, ma di fatto rischiose e ancora limitate dalle possibilità tecnologiche attuali, nonché dalla scarsa consuetudine degli utenti nell'interazione con sistemi automatici. Far ripetere l'enunciato all'utente può creare tuttavia altri problemi: in particolar modo, la ripetizione dell'enunciato può rimanere problematica per il sistema, se l'errore precedente non era legato a un'errata decodifica del riconoscitore vocale ma, per esempio, a un termine assente nel vocabolario; inoltre, è stata verificata la tendenza degli utenti a iperarticolare la ripetizione, fenomeno che riduce di molto le prestazioni del riconoscitore vocale, generalmente addestrato sul parlato spontaneo (cfr. 5.3.1).

Questo genere di problemi, come vedremo nell'analisi delle riformulazioni, sono legati all'idea che ha il parlante delle capacità di comprensione della macchina e di ciò che può aver causato l'errore: in questo è favorito dal fatto che, generalmente, gli output che segnalano un'incomprensione chiedono la ripetizione dell'enunciato senza dare alcun punto di riferimento all'utente sul problema che è occorso. Larsson 2003 propone una varietà di output di gestione degli errori legata a differenti livelli comunicativi: il sistema può così segnalare se non ha identificato alcun segnale

vocale, se lo ha identificato ma non riesce a decodificarlo (ASR), se non ne ha compreso il significato, se non ha inteso le intenzioni, se una determinata richiesta non è prevista per l'applicazione. D'altra parte, molti lavori sull'usabilità dei sistemi di dialogo criticano un tale approccio poiché spiegare un errore non aiuta necessariamente il parlante, mentre consigliano di usare le richieste di ripetizione come occasione per indirizzare l'utente su quello che può dire nel turno successivo (Balentine – Morgan 1999).

6.4.4 Simulazione degli errori

Come già descritto (cfr. 5.6.3.5), il comportamento del wizard è stato definito da un protocollo, redatto *a priori* sulla base dell'architettura del futuro sistema. Ai fini della simulazione, il wizard si è servito di un insieme di enunciati (o output) fissi organizzati in alberi di dialogo, attenendosi al protocollo predefinito, pur con la possibilità di intervenire localmente sull'interazione nel caso di situazioni pertinenti per il compito ma non previste nella fase di progettazione. Al contrario, non sono state poste particolari restrizioni sul versante della comprensione: il parlante non è stato quindi condizionato da limitazioni legate a uno specifico modulo di comprensione del linguaggio o da una specifica gestione dell'iniziativa.

Come illustrato da Riek 2012 in una rassegna sugli studi empirici condotti con la tecnica del Mago di Oz, è usuale che i moduli di comprensione del linguaggio naturale e di gestione del dialogo siano simulati senza prevedere particolari limitazioni per lo sperimentatore: sia per limitazioni strutturali (è difficile individuare in tempo reale quello che un sistema potrebbe non capire di quanto detto dall'utente), sia perché, generalmente, proprio questi moduli devono essere implementati sulla base dei dati ottenuti durante la raccolta. Si corre tuttavia il rischio che le interazioni tra soggetti e sistema simulato non presentino le situazioni di errore che occorrono più di frequente nel dialogo uomo-macchina.

Le difficoltà che normalmente intercorrono nell'interazione con un sistema automatico possono essere simulate in vari modi: interponendo il riconoscitore vocale tra wizard e il segnale audio del parlante, di modo che il wizard abbia accesso solo a ciò che viene trascritto dal riconoscitore (che, in tal caso, deve essere lo stesso del sistema); disturbando il segnale audio, affinché sia più difficile per il wizard comprendere ciò che viene detto dal parlante; simulando anche gli errori nel corso

dell'interazione (per esempio, lasciando il loro uso a discrezione del wizard o stabilendo prima quando devono essere usati durante l'interazione). Frapporre elementi di disturbo tra wizard e soggetto ha lo scopo di simulare in maniera realistica le difficoltà di comprensione del sistema, de-umanizzando lo sperimentatore che svolge il ruolo del sistema; tuttavia, è difficilmente eliminabile la componente umana: per quanto la comprensione completa di ciò che viene detto al sistema sia compromessa, un essere umano può facilmente recuperare le informazioni mancanti conoscendo il contesto e gli obiettivi. Invece, stabilire aprioristicamente gli errori, più che puntare a una simulazione realistica dei problemi che possono intercorrere nella comprensione, punta a ridurre la componente umana del wizard, riducendone l'interpretazione personale e obbligandolo a rispettare determinate regole definite in fase di progettazione.

6.4.4.1 Simulazione degli errori nel corpus Speaky.

Nella simulazione del sistema abbiamo scelto di inserire nell'architettura alcuni output predefiniti di non comprensione ed errata acquisizione. Avendo previsto per il futuro sistema un controllo robusto dell'acquisizione, abbiamo simulato le richieste di conferma esplicite e le richieste di ripetizione, generalmente usate per i livelli più bassi di *confidence*.

6.4.4.1.1 Richieste di ripetizione: per la simulazione delle situazioni di non comprensione da parte del sistema sono stati scritti tre output sostanzialmente equivalenti (come verificato da un'indagine preliminare del *corpus*)⁶¹: “non ho capito, puoi ripetere?”, “non ho capito, potresti ridirmelo?”, “non riesco a comprendere, puoi ripetere?”. Ogni compito prevedeva una richiesta di ripetizione, collocata, di preferenza, nella fase di apertura di un compito o di un sottocompito (ovvero, nelle fasi più complesse da gestire per il parlante).

WOZ: non serve la ricetta del medico

AN26: bene, me lo puoi ordinare, al più presto?

WOZ: non riesco a comprendere, puoi ripetere?

⁶¹ Non sembrano esserci relazioni particolari tra le richieste di ripetere la frase da parte del sistema e la preferenza per una strategia di correzione e ripetizione dell'enunciato non compreso dal sistema. La proposta di Larsson 2003 di differenziare gli output a seconda del livello comunicativo in cui si colloca l'incomprensione andrebbe verificata proprio sull'uso da parte dei parlanti.

AN26: puoi ordinare il farmaco in farmacia?

WOZ: sì, vuoi pagare ora con carta di credito o pagare alla consegna?

Le richieste di ripetizione sono state usate anche in casi decisi in tempo reale dal Mago, nei casi identificabili come impossibili da gestire da parte del futuro sistema, nel caso di richieste fuori dominio, non attinenti al compito, nel caso di turni con eccessive esitazioni e false partenze (per evitare un comportamento troppo “naturale” da parte del soggetto) e in situazioni meno prevedibili, in cui il wizard ha richiesto la riperitizione per guadagnare tempo quando vi è stata la necessità di usare il *real-time* (cfr. capitolo 5).

6.4.4.1.2 Richieste di conferma esplicite: a differenza della simulazione della non comprensione, che può avvenire in qualsiasi momento dell'interazione, l'uso delle richieste di conferma, essendo determinato da un'errata acquisizione da parte del sistema, è legato al contesto locale e a ciò che è stato detto durante l'interazione. In ogni compito è stata prevista un'errata acquisizione, per un totale di 40 output. Nella maggior parte degli output viene chiesto al parlante di confermare un'informazione, che può essere una singola variabile, come nell'esempio seguente:

WOZ: ciao, come posso aiutarti?

NV09: ho bisogno di andare a milano, il<cc> ventisette aprile, potresti vedere
che treni ci sono la mattina del ventisette

WOZ: vuoi un treno per merano?

NV09: no per milano

Oppure, più raramente, può essere richiesta nello stesso turno la conferma di più dati acquisiti:

NV09: possiamo impostare la temperatura, di ventitré gradi nella mia camera
da letto?

WOZ: vuoi impostare la temperatura del salotto a venticinque gradi?

NV09: no, voglio impostare la temperatura di ventitré gradi, in camera da
letto

In alcuni casi viene chiesta conferma dell'azione da compiere:

WOZ: ciao, come posso aiutarti?

AN14: <ee> volevo sapere l'elenco dei film registrati

WOZ: vuoi registrare un film?

AN14: no, volevo sapere <ee> i film registrati di ieri

Nella simulazione degli errori abbiamo anche inserito due output che non sono richieste di conferma, bensì feedback di completamento con un'errata acquisizione (situazione che naturalmente può presentarsi durante l'interazione quando un errore è sfuggito ai meccanismi di controllo e correzione): un output simula l'errata esecuzione di un'azione, l'altro segnala l'impossibilità di compiere un'azione, dopo un'errata acquisizione:

WOZ: va bene, comunico al servizio taxi che stamattina non devono passare a prenderti

NV03: ma, da domani, conferma servizio

WOZ: va bene, il taxi non passerà per tutta la settimana

NV03: no, da, domani, voglio, che il taxi, passi

WOZ: ciao, ti serve qualcosa?

NV04: puoi annullare l'appuntamento dal dentista, <ee> il 5 dicembre alle quattordici?

WOZ: non ho nessun appuntamento registrato per il 20 dicembre

NV04: no, il, cinque dicembre, alle quattordici

Complessivamente sono state usate 687 richieste di ripetizione e 523 richieste di conferma con un'acquisizione errata, che corrispondono ad altrettante situazioni in cui l'utente ha dovuto confrontarsi con errori da parte del sistema.

6.4.5. Situazioni di mancata comprensione

6.4.5.1 Sequenza di riparazione.

La sequenza dei turni che si presenta durante le situazioni di mancata comprensione è fissa; l'architettura prevede infatti una risoluzione della mancata comprensione in due turni: dopo la segnalazione della mancata comprensione, il parlante può ripetere la propria richiesta e, se questa viene compresa dal sistema, continuare l'interazione:

WOZ: ciao, in cosa posso esserti utile?

AN03: programma il sistema di irri* irrigazione sul mio terrazzo

WOZ: non ho capito, puoi ripetere?

AN03: programma, il sistema di irrigazione

Va tuttavia notato che, proprio durante la richiesta di ripetizione si accorciano i tempi di presa del turno da parte dell'utente (cfr. 6.3.3.4.), portando così a sovrapposizioni (molto rare) nel resto del *corpus* (20 casi); in particolare, ciò avviene dopo la pausa breve che, in tutti e tre gli output, separa l'asserzione sulla mancata comprensione e la richiesta di ripetizione. Nella maggior parte dei casi (16) il parlante si interrompe immediatamente, riformulando l'enunciato una volta che il sistema ha finito di parlare:

WOZ: vuoi pagare ora con carta di credito o pagare alla consegna?

AN03: pagarlo alla consegna

WOZ: non ho capito, # puoi ripetere? #

AN03: # pagarlo #. pagarlo alla consegna

NV02: potrei spegnerle

WOZ: non ho # capito #, # puoi # ripetere?

NV02: #vorrei # # vorrei # vorrei chiuderle, e inserire
l'antifurto

WOZ: ho chiuso le finestre del salotto e della cucina

Sono invece molto rari i casi (4) in cui il parlante continua la riformulazione, senza ripetere la parte del turno sovrapposta a quella del sistema:

AN21: sì da spengere, sono da spengere
WOZ: non riesco a comprendere, # puoi ripetere? #
AN21: # sono da spegnere #
WOZ: ok, tutte le luci di casa sono spente

L'impossibilità di simulare il *barge-in* (cfr. 5.6.3.5) ha senz'altro ridotto di molto la presenza di sovrapposizioni nel *corpus*, rendendo poco funzionale la sovrapposizione al turno del sistema come strategia conversazionale durante la risoluzione di un errore.⁶²

6.4.5.2 Ripetizioni e riformulazioni

Quando viene simulata una mancata comprensione da parte del sistema, al parlante viene chiesto di ripetere l'enunciato. Dopo tale richiesta, l'utente può ripetere l'enunciato precedente o riformularlo cercando di renderlo intellegibile al sistema: sulla base delle esperienze precedenti con computer (e relative interfacce) e con interfacce vocali (Bell – Gustafson 2003), e in consonanza con l'esperienza acquisita con il sistema. In generale, la situazione di errore e la sua risoluzione portano il parlante a confrontarsi con le capacità linguistiche del sistema, spesso senza avere un feedback di ciò che il sistema ha effettivamente inteso⁶³.

Come vedremo a proposito delle ripetizioni, per la lingua inglese sono state condotte varie indagine relativamente alla risoluzione degli errori da parte del parlante nell'interazione uomo-macchina e alle caratteristiche degli enunciati successivi all'errore. Le variazioni che sono state approfondite maggiormente sono quelle di carattere fonetico e prosodico. Infatti, la gestione degli errori, come abbiamo già accennato nel paragrafo 3, è affidata principalmente ai risultati del

⁶² Le 20 sovrapposizioni sono distribuite tra 17 soggetti: solo 3 soggetti si sovrappongono al turno del sistema più di una volta. Va tuttavia precisato che tutte le sovrapposizioni si presentano quando il soggetto si è già confrontato con situazioni di mancata comprensione e con il relativo output; non sono quindi dovute a un'errata valutazione del turno del sistema (cfr. 6.3.3.6.1).

⁶³ Non tutti i sistemi non danno feedback su ciò che hanno compreso dell'enunciato del parlante, anche quando viene chiesta la ripetizione: ad esempio Siri mostra la decodifica del riconoscitore vocale, dando al parlante la possibilità di capire se l'errore è dovuto a un'errata decodifica, oppure a un termine che il riconoscitore trascrive ma non fa parte del modello di comprensione del sistema. Il feedback su ciò che è stato realmente decodificato dal sistema condiziona naturalmente la risoluzione dell'errore, fornendo al parlante maggiori informazioni per la risoluzione; d'altra parte, l'opzione è comoda solo per sistemi di dialogo che coinvolgano anche la lettura e il feedback visivo, come uno smartphone.

riconoscitore vocale: incomprensioni o mancate acquisizioni sono individuate dal sistema principalmente tramite la definizione della *confidence* per ogni input. A rinforzare l'interesse per l'aspetto acustico si aggiunge il generale peggioramento dei risultati dell'acquisizione successivamente a un errore, causato da fenomeni di iperarticolazione presenti nei turni successivi all'errore: il parlante, dopo una mancata acquisizione, tende a pronunciare con maggiore chiarezza il proprio enunciato, ad allungare e aumentare le pause tra una parola e l'altra, a rallentare la velocità d'eloquio, ad aumentare l'intensità della produzione vocale. Questi fenomeni, che nelle conversazioni quotidiane migliorano generalmente la comprensione dell'interlocutore, peggiorano la *performance* del riconoscitore vocale.⁶⁴

Viste la cause degli errori (principalmente legate al riconoscitore vocale) e vista la semplicità di lessico e strutture sintattiche normalmente coinvolte nell'interazione con sistemi commerciali a iniziativa ridotta per l'utente, le alterazioni relative ad altri livelli linguistici sono state molto meno indagate in letteratura.

I pochi studi presenti sulle modificazioni lessicali e sintattiche sono principalmente di tipo quantitativo: Gustafson – Bell 2000 rilevano che il principale mutamento nelle situazioni di mancata comprensione è la modifica di uno degli elementi lessicali dell'enunciato precedente, seguita dall'inserzione/eliminazione di una proposizione o parola e da un cambiamento dell'ordine sintattico; anche Oviatt et al. 1998 segnalano, tra le modificazioni che possono occorrere dopo un errore, la tendenza a variare il lessico. Parzialmente pertinente, e incentrato su altri piani linguistici (pragmatica, sintassi e lessico), è il lavoro di Krahmer (Krahmer et al. 1999, 2001), che ha analizzato il comportamento successivo a un errore allo scopo di elaborare nuove metriche per supportare la risoluzione di un'incomprensione dopo richieste di conferma esplicite e implicite, individuando così alcune caratteristiche dei turni successi a un errore: maggiore lunghezza, sintassi marcata, segnali di disconferma, mancanza di nuove informazioni, ripetizione delle parole precedenti.

6.4.5.3 Ripetizioni e riformulazioni nel *corpus* Speaky

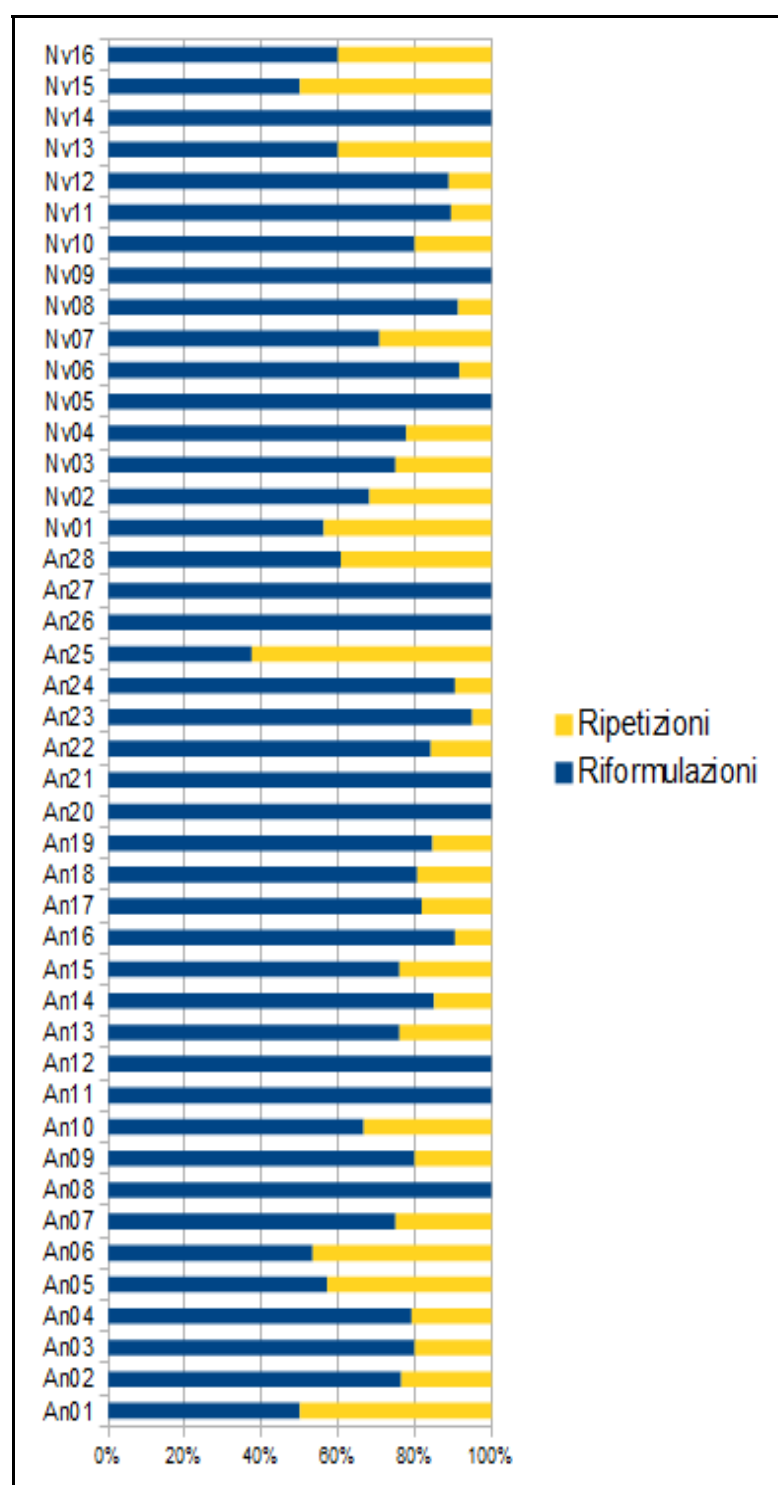
Sia la ripetizione sia la riformulazione fanno parte dello stesso *continuum* delle iterazioni: la differenza, infatti, è tutt'altro che netta: «many repetitions are not

⁶⁴ Bisogna tuttavia ricordare che tali parametri acustici sono stati principalmente analizzati relativamente alle ripetizioni, che, stando alle ricerche condotte su vari corpora, sono minoritarie rispetto alle riformulazioni: Boyce 2008, Gustafson – Bell 2000.

exact, and many reformulations include repetitions of lexical items or grammatical structures from the source utterance» (Bamford 2000). Nell'analisi del *corpus* la distinzione è stata fatta sulla base di alcuni fattori: sono stati considerati ripetuti gli enunciati che presentano i medesimi elementi morfologici e lessicali, e la medesima struttura sintattica. La variazione dovuta agli stretti tempi di programmazione del parlato (esitazioni, pause piene, segnali discorsivi e false partenze) non è stata tenuta in considerazione al fine della divisione tra le due diverse tipologie, in quanto non sufficientemente rilevante per poter individuare un'intenzionalità di variazione nella ripetizione dell'enunciato. La riformulazione, invece, può avvenire su vari livelli (morfologico, sintattico, lessicale e pragmatico), fino alla sostituzione completa di ogni elemento della frase originaria, la quale resta affine al nuovo enunciato perché ha le medesime intenzioni pragmatiche. Per quel che riguarda le riformulazioni rimane problematica la distanza semantica che può intercorrere tra l'enunciato originario e quello riformulato, problema che “coinvolge preliminarmente quello del significato linguistico” (Sornicola 1981: 260). In una prospettiva pragmatica e conversazionale, è necessario distinguere tra riformulazioni linguistiche (i due enunciati sono in relazione parafrastica per ogni locutore, in ogni situazione) e riformulazioni pragmatiche (l'enunciato riformulato presenta la medesima intenzione e rinvia agli stessi fatti dell'enunciato originario, cfr. Mortara Garavelli 1979). Nel presente lavoro, fortemente incentrato sul parlato e sul dialogo, le riformulazioni sono da intendersi come “parafrasi pragmatiche”.

Nel *corpus* *Speaky* la strategia preferita da parte dei parlanti successivamente a una situazione di mancata acquisizione è la riformulazione di almeno uno dei componenti dell'enunciato: su 687 casi di mancata acquisizione abbiamo 138 casi di ripetizione (21%) e 529 riformulazioni (79%). Come nota Fischer 2006 le scelte fatte dopo errore da parte del sistema sono strettamente collegate all'idea che ha il parlante della macchina e da quanto (e come) provi ad andarle incontro. Nel nostro caso la riformulazione è la strategia preferita da 41 parlanti su 44; solo un parlante (anziano, donna) preferisce la ripetizione dell'enunciato rispetto alla sua riformulazione, mentre per gli altri due (anziano, donna; non vedente, donna) ripetizione e riformulazione sono usate nella stessa misura.⁶⁵

⁶⁵ Al contrario, il *corpus* C-ORAL-ROM di parlato uomo-macchina, presentava, anche laddove l'iniziativa mista era ampiamente sfruttata, una decisa preferenza per le ripetizioni da parte di tutti i parlanti.



Graf. 21: Preferenza tra ripetizione e riformulazione per ogni parlante

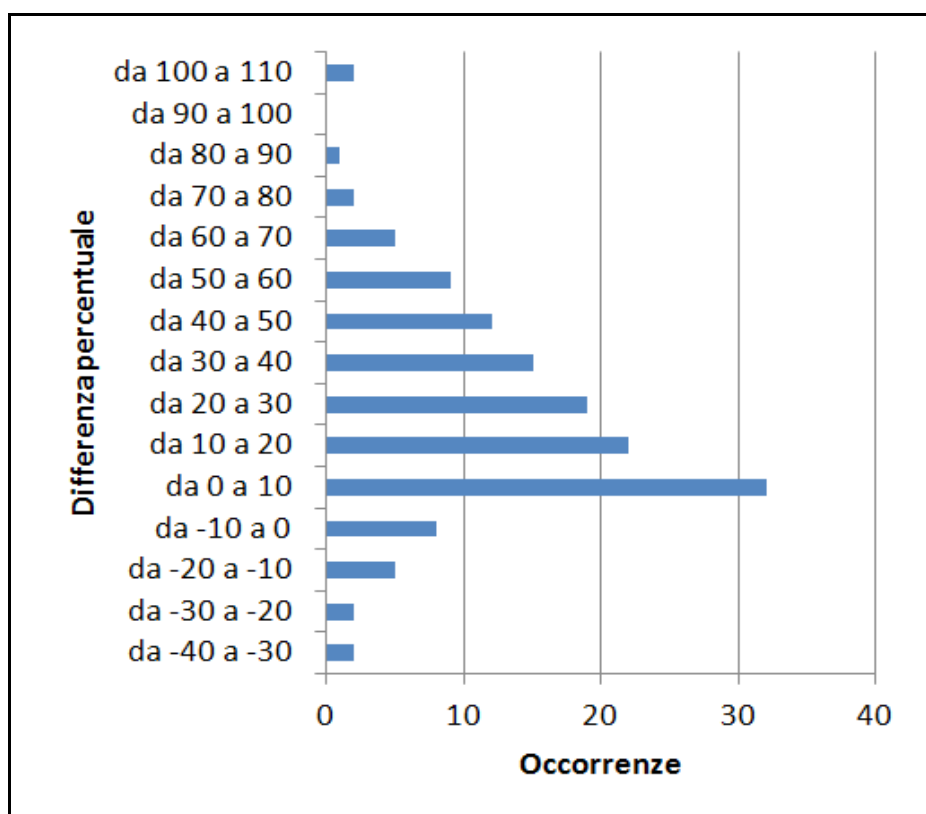
6.4.5.3.1 Ripetizioni

Per quanto vi sia una completa invarianza dal punto di vista sintattico e lessicale, il parlante, nella ripetizione, può variare l'esecuzione dell'enunciato a livello acustico e prosodico, come già evidenziato in numerosi studi sulle situazioni di errore (Oviatt et al. 1998, Levow 1998): la variazione prosodica principale è l'iperarticolazione dell'enunciato, a cui corrisponde un rallentamento della velocità d'eloquio, una maggiore intensità acustica, la presenza di pause più lunghe tra una parola e l'altra (e talvolta tra le sillabe di una parola). Questi fenomeni che possono occorrere durante la ripetizione dell'enunciato mettono spesso in difficoltà i riconoscitori vocali, generalmente addestrati e tarati sul parlato spontaneo.

Pur non avendo preso in esame l'aspetto acustico e prosodico nel lavoro, la presenza nel *corpus* di fenomeni di iperarticolazione durante la ripetizione di un enunciato è stata verificata tramite il parametro della differenza di durata tra gli enunciati ripetuti, calcolata in percentuale. La durata è stata calcolata a partire dall'inizio dell'articolazione dell'enunciato fino al suo completamento, non considerando le pause precedenti e le pause successive, prima che il sistema prendesse il turno.

Va precisato che un enunciato iperarticolato e ripetuto non è automaticamente più lungo dell'originario⁶⁶; al contrario, un aumento della durata dell'esecuzione corrisponde probabilmente alla presenza di parlato iperarticolato. Complessivamente le ripetizioni sono più lunghe degli enunciati originali del 21%.

⁶⁶ Ad esempio, la differenza di durata può essere cancellata dalla presenza di una lunga pausa nell'enunciato originario; tuttavia, nella misurazione non sono stati riscontrati casi in cui un enunciato chiaramente iperarticolato fosse meno lungo del precedente.

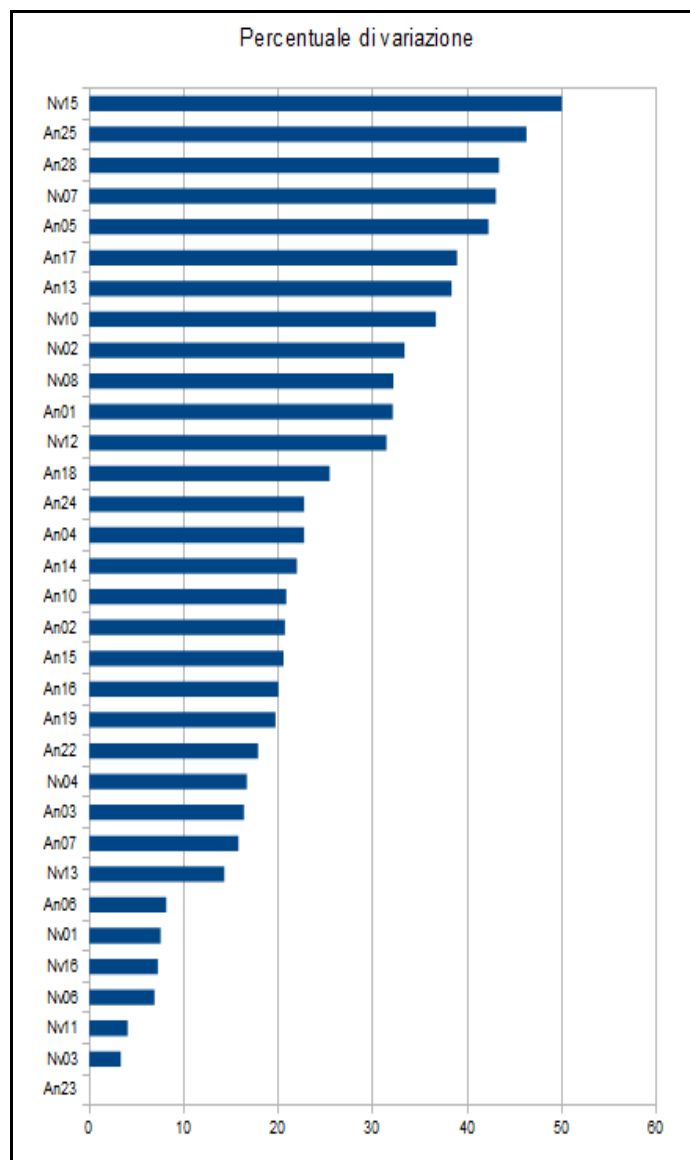


Graf. 22: Variazione di durata dell'input dopo una richiesta di ripetizione

Come si può vedere dal grafico 23, sono presenti enunciati ripetuti con minore durata dei precedenti: questo avviene generalmente quando nel turno originario sono presenti pause molto lunghe, dovute alla pianificazione dell'enunciato, che nella ripetizione vengono soppresse, dato che non c'è bisogno di una nuova pianificazione. Contando solo le ripetizioni senza alcuna variazione (o con variazione minima) e quelle più lente, si arriva a un rallentamento medio della velocità di eloquio del 26% (fino a 32% se vengono tolte dalla media le ripetizioni dello stessa durata dell'enunciato originario).

Anche relativamente all'iperarticolazione è molto accentuata la variazione interpersonale, il che rende difficile sintetizzare una strategia condivisa di ripetizione tra tutti i parlanti. La variazione tra enunciati originali ed enunciati ripetuti, distribuita tra i parlanti (escludendo i 10 utenti che durante l'interazione hanno solo riformulato l'enunciato), mostra una forbice molto larga: si va da una variazione praticamente impercettibile (che non si può ascrivere all'iperarticolazione) a un'iperarticolazione molto accentuata che può costituire a tutti gli effetti una strategia

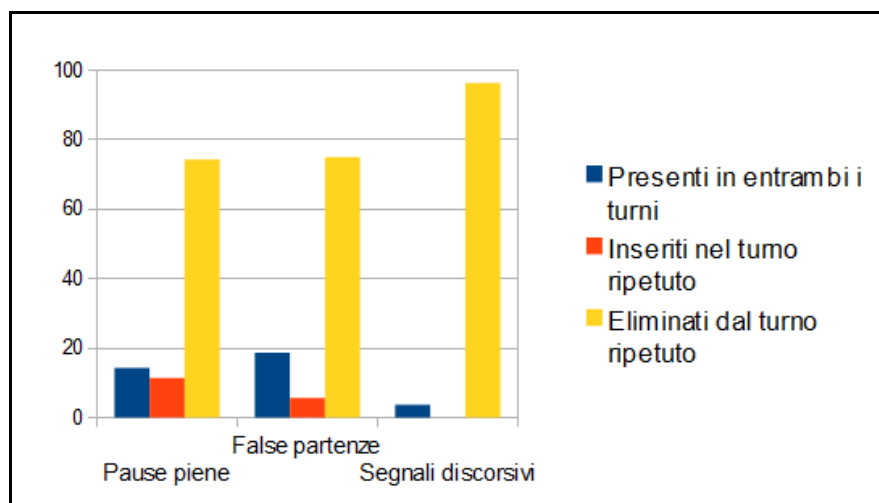
di risoluzione dell'errore. In particolare, incrociando questi dati con la preferenza tra ripetizioni e riformulazioni, possiamo ricavare che i tre soggetti che preferiscono evitare la riformulazione, presentano un'alta percentuale di variazione della durata dell'enunciato: rispettivamente hanno infatti una variazione del 50%, del 46% e del 32%:



Graf. 23: Percentuale di variazione di durata dell'input per ogni soggetto dopo una richiesta di ripetizione

Come già detto, le ripetizioni che hanno minore durata rispetto dell'enunciato originario sono spesso dovute alla minore presenza di pause ed esitazioni, connesse alla pianificazione dell'enunciato; l'errore permette infatti al parlante di ripetere un turno già precedentemente programmato. Le ripetizioni sono così caratterizzate da

una maggiore “pulizia” e programmazione: diminuiscono infatti pause piene, false partenze e segnali discorsivi.



Graf. 24: Presenza (in percentuale) di fenomeni del parlato nel turno ripetuto

Se ascriviamo i comportamenti dopo un errore all’interpretazione che il parlante dà relativamente alle cause dell’errore occorso e alle capacità di comprensione del sistema, possiamo cogliere nelle ripetizioni la strategia per risolvere un errore che è stato identificato come problema del canale (il comando non è arrivato) o del riconoscimento (il comando non è stato sentito distintamente). L’iperarticolazione che contraddistingue gli enunciati ripetuti, in aggiunta alla maggiore “pulizia” e programmazione, concorrono nell’identificare il problema verificatosi durante il turno come un problema di decodifica del comando da parte del sistema. Le ripetizioni con variazioni acustiche e prosodiche minime, invece, sembrano legarsi maggiormente all’individuazione di un problema dipendente dal canale di comunicazione più che dal riconoscimento.

6.4.5.3.2 Riformulazioni

Possiamo parlare di riformulazione quando il parlante interviene modificando lessico e sintassi del nuovo enunciato. Nel *corpus* sono molto più frequenti i casi in cui vi è un’alterazione della struttura e del lessico dell’enunciato rispetto alla sua riproposizione: in questi casi il parlante può intervenire su varie componenti dell’enunciato, a seconda delle diverse interpretazioni della situazione di errore e

delle cause che hanno portato alla non comprensione.

6.4.5.3.2.1 Riformulazioni dello svolgimento del compito

Un primo livello d'intervento è di natura pragmatica e riguarda la programmazione dell'esecuzione del compito. Tale variazione non può essere presente nell'interazione con architetture a stati finiti e a iniziativa del sistema, poiché è il sistema stesso a indirizzare l'esecuzione del compito, come una qualsiasi interfaccia grafica; nei sistemi a iniziativa mista, a meno di particolari indicazioni negli output del sistema su ciò che va detto in quel turno⁶⁷, è lasciata la libertà al parlante non solo di formulare le richieste come desidera, ma anche di organizzare lo svolgimento dell'attività.

L'utente non sa in quale modo i dati e le informazioni che dà al sistema vengano gestiti dallo stesso per eseguire il compito; per questo la programmazione del compito e delle informazioni da fornire al sistema rientrano tra le possibili cause di un'incomprensione, e la variazione del carico informativo può essere una strategia per risolvere l'incomprensione.

L'enunciato riformulato può avere una densità informativa ridotta, per diminuire il carico di lavoro del sistema nel singolo turno e distribuirlo su più turni:

WOZ: ti serve altro?

AN13: sì, prenota il taxi, alle <mm> venti e quindici

WOZ: non ho capito, puoi ripetere?

AN13: prenotare il taxi

WOZ: per che ora?

AN13: venti e quindici

WOZ: ecco i ristoranti con cucina romana aperti fino a tardi vicino casa tua: i tre moschettieri, da zia maria, tutto alla brace. vuoi fare una

⁶⁷ Indicazioni per forza generiche, via via che i compiti e i domini da gestire sono più complessi. Si veda a tal proposito l'output di esordio del sistema del corpus C-ORAL-ROM (cfr. 4.2.2.1), in cui vengono proposte entrambe le strategie ma non viene detto al parlante tutto quello che può dire; lo stesso vale per il famoso output iniziale "How may I help you?" (Gorin et al., 1997), in cui non viene data alcuna indicazione al parlante. Le ragioni sono legate all'usabilità minima del sistema: un *output* iniziale troppo lungo può dare una pessima impressione al parlante (basti pensare alla scomodità delle interfacce DTMF in cui vengono elencate tutte le opzioni relative a ogni numero della tastiera); oltre al fatto che un sistema a iniziativa mista si propone proprio, come obiettivo, di non dover dare istruzioni all'utente e di interpretare la sua pianificazione del compito.

prenotazione?

NV11: sì

WOZ: va bene, per quante persone?

NV11: per due, al ristorante, da zia maria

WOZ: non ho capito, puoi ripetere?

NV11: per due

WOZ: va bene, per che ora?

NV11: per le dieci di sera

WOZ: va bene, in quale ristorante?

NV11: da zia maria

L'omissione può anche riguardare parti del discorso che non hanno rilevanza per il sistema e che vengono cassate dall'utente in favore di quelle con maggiore forza illocutoria:

WOZ: ciao, in cosa posso esserti utile?

AN08: ho bisogno di fare un esame del sangue, e avrei bisogno di recuperare la vecchia prescrizione, che ho fatto a giugno

WOZ: non ho capito, puoi ripetere?

AN08: devo recuperare la vecchia prescrizione delle analisi del sangue che ho, che ho fatto a giugno

Altrimenti, il carico informativo può essere aumentato, con nuovi dettagli, allo scopo di specificare ed esplicitare il contesto dell'azione:

WOZ: sono rimaste ancora accese le luci della cucina e del salotto

AN13: allora devi spegnere le luci

WOZ: non riesco a comprendere, puoi ripetere?

AN13: devi spegnere le luci in tutte le stanze

WOZ: ciao, ti serve qualcosa?

AN16: vorrei delle informazioni meteo

WOZ: non ho capito, potresti ridirmelo?

AN16: sì, vorrei informazioni meteorologiche sulla / su roma, per i prossimi

giorni

L'utente può arrivare a cambiare completamente l'obiettivo del turno: in tal caso si può parlare di riprogrammazione più che di riformulazione,⁶⁸ venendo a mancare il parallelismo pragmatico necessario tra i due enunciati per poter identificare nel nuovo una nuova formulazione dell'originario:

WOZ: ecco i ristoranti aperti fino a tardi vicino casa tua: i tre moschettieri, da zia maria, tutto alla brace ti # serve # altro?

AN02: # ma # eh vorrei sapere il
<ee> il tipo di cucina

WOZ: non ho capito, puoi ripetere?

AN02: non importa <ee>. scegliamo <ee> da zia maria

Le variazioni del carico informativo, oltre a incidere sulla struttura conversazionale (implicando una diverso numero di turni a seconda della variazione), incidono sulla struttura sintattica, portando infatti a espansioni o riduzioni, sia a livello frasale che a livello interfrasale. A fronte della preferenza per la riformulazione linguistica, non si riscontra la stessa preferenza nel variare la programmazione dell'esecuzione del compito; infatti, delle 539 riformulazioni, l'82% ha lo stesso contenuto di informazioni per il sistema, mentre solo il 9% presenta una riduzione, il 5% un aumento, il 4% un variazione totale rispetto all'enunciato precedente (quest'ultimo gruppo è stato escluso dal resto dell'analisi non essendo

⁶⁸ Vale la pena ricordare, i vizi strutturali dell'ambiente sperimentale e della tecnica del Mago di Oz che incidono sulla presenza di cambi di intenzioni durante l'interazione. Infatti, per quanto la tecnica consenta di simulare un sistema, il contesto di interazione è diverso da quello che si viene a stabilire con utenti reali, che interagiscono con il sistema con motivazioni e necessità reali; motivazioni che non possono essere presenti nella stessa misura in un ambiente sperimentale e portano talvolta all'abbandono di un obiettivo. Tuttavia, il diverso coinvolgimento tra utenti reali e soggetti sperimentali non spiega interamente le riformulazioni del compito, presenti anche in *corpora* d'interazioni reali e spesso dovute a un cambio di obiettivo limitato al singolo turno e a una nuova pianificazione dello svolgimento, come nell'esempio seguente, in cui il soggetto prima chiede conferma del numero di cellulare, e poi, dopo l'incomprensione del sistema, decide di inserirlo comunque:

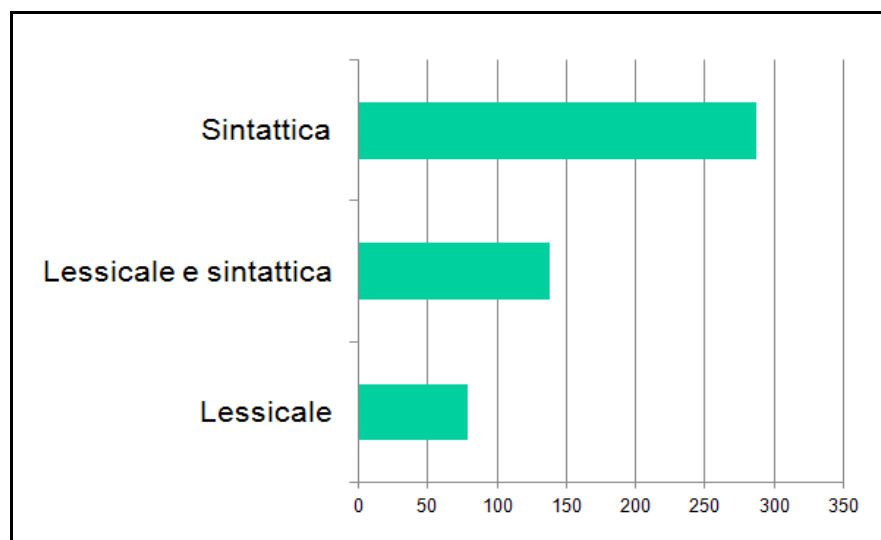
WOZ: va bene, la deviazione delle chiamate sul cellulare si attiverà a mezzanotte.
ti serve altro?
AN: eh sapere il numero del cellulare
WOZ: non ho capito, puoi ripetere?
AN: sì, ti do il numero del cellulare

possibile un confronto linguistico tra gli enunciati). Andando a sommare questi dati a quelli delle ripetizioni, che non presentano variazioni, possiamo considerare come largamente predominante nel *corpus* Speaky la scelta da parte dei parlanti di non riprogrammare la risoluzione del compito nelle situazioni di non comprensione: nell'86% dei casi il parlante lascia invariato il contenuto significativo dell'enunciato, variandone, invece, la composizione sintattica e lessicale, con conseguenti risvolti pragmatici e testuali. In questo i risultati differiscono in parte da lavori in cui è stato analizzata la riprogrammazione e il cambio pragmatico: in Bohus 2007 i cambi pragmatici sono nel 31% del *corpus* analizzato, mentre in Boyce 2008 sono complessivamente del 24%.⁶⁹

6.4.5.3.2 Riformulazioni sintattiche e lessicali

Con le ripetizioni (e le eventuali variazioni acustiche e prosodiche) il parlante può risolvere problemi del canale (il sistema non ha sentito) o della percezione/trascrizione (il sistema non ha decodificato le parole). Con la riprogrammazione il parlante prova a risolvere i deficit di esecuzione del compito (il sistema non è in grado di svolgere il compito in questo modo). Tuttavia, nella maggior parte dei casi, i parlanti attribuiscono la non comprensione a problemi linguistici, relativi ai termini usati, alla presentazione e formulazione delle richieste, e al *common ground*. La maggior parte delle variazioni tra l'enunciato originario e la sua riformulazione avviene infatti su un piano prettamente linguistico: in particolar modo lessicale e sintattico. L'enunciato riformulato può così presentare lo stesso costrutto dell'originario con elementi lessicali sostituiti (sinonimi, iponimi, iperonimi ecc.), ridotti o espansi; allo stesso modo può essere modificata la composizione sintattica dell'enunciato (spesso per ragioni pragmatiche legate alla cortesia), ridotta o espansa, a seconda di ciò che il parlante pensa riguardo alla conoscenza cotestuale del sistema e alla memoria di ciò che è stato detto precedentemente; infine, la variazione può toccare entrambi i piani contemporaneamente. La maggior parte delle riformulazioni avviene sul piano sintattico e coinvolge anche il lessico, mentre la sola modificazione lessicale priva di effetti sulla sintassi è meno frequente.

⁶⁹ Il *corpus* RoomLine analizzato da Bohus 2007 è stato raccolto in parte con il vero sistema, in parte in un ambiente Wizard of Oz: bisogna tenere presente che gli output usati per segnalare l'incomprensione afferiscono a dieci diverse strategie di risoluzione dell'errore. Boyce 2008 usa un corpus, di tipo HMIHY, raccolto con la tecnica del Mago di Oz: l'analisi include anche l'aggiunta o l'omissione di informazione, che, rispettivamente coprono il 12% e l'8%, a cui si aggiunge il 4% di enunciati con un cambio di obiettivo nel singolo turno.



Graf. 25: Distribuzione delle riformulazioni secondo il tipo di variazione

6.4.5.3.2.2.1 Riformulazione lessicale. Se il sistema non comprende un enunciato, una delle ipotesi che può formulare il parlante è che uno dei termini usati non sia stato compreso dalla macchina perché non presente nel suo vocabolario. Nel 41% degli enunciati riformulati (217) i soggetti si affidano alla *variatio* lessicale nel tentativo di risolvere l'errore. La strategia predominante di variazione (84%) è la sostituzione di uno degli elementi lessicali con un altro elemento in rapporto di sinonimia, iponimia o iperonimia:

WOZ: ciao, come posso aiutarti?

AN12: dovresti <ee> chiudere le tapparelle, in salotto

WOZ: non ho capito, puoi ripetere?

AN12: <ee> dovresti, abbassare, le tapparelle, in salotto

WOZ: va bene, la sveglia è stata impostata alle 7 di mattina tutti i giorni della settimana ad esclusione del weekend

NV06: <ee> posso chiederti di sintonizzarlo, / di sintonizzare la sveglia su un canale di notizie?

WOZ: non ho capito, puoi ripetere?

NV06: se puoi, per cortesia, sintonizzare la sveglia / la radiosveglia su un canale di notizie

WOZ: per il ritorno ce n'è uno che parte alle nove e trenta da milano centrale.
vuoi altre informazioni?

NV13: voglio un altro treno

WOZ: non ho capito, puoi ripetere??

NV13: voglio un'altra soluzione

Un altro meccanismo frequente di sostituzione lessicale è l'esplicitazione di un elemento anaforico, di cui rimandiamo l'approfondimento al paragrafo dedicato alle ragioni coesive di alcuni meccanismi di variazione (cfr. 6.4.5.3.2.2.3), essendo una sostituzione che afferisce più a ipotesi sulla memoria discorsiva del sistema che a *deficit* del vocabolario, come si può vedere nel seguente esempio:

WOZ: va bene, ho disdetto l'appuntamento del 5 dicembre alle 14

AN15: bisogna prenderne uno nuovo per il giorno sette dicembre alle quindici

WOZ: non ho capito, potresti ridirmelo?

AN15: bisogna prendere un nuovo appuntamento per il giorno sette dicembre
alle quindici

Meno frequenti (18%) sono le espansioni del sintagma nominale, che spesso coincidono con una maggiore formalità, tramite l'esplicitazione di parti del sintagma normalmente ellittiche nell'uso comune, si veda ad esempio:

WOZ: ti serve altro?

NV14: pr* prenotami per due <ee> tutto alla brace

WOZ: non ho capito, puoi ripetere?

NV14: ok prenotami per due persone nel ristorante tutto alla brace

WOZ: ciao, ti serve qualcosa?

NV16: sì, controlla, credito, telefono

WOZ: non ho capito, potresti ridirmelo?

NV16: dimmi, il credito, del telefono cellulare

Oltre alla maggiore formalità, l'espansione del sintagma nominale conduce

generalmente a una maggiore specificazione della richiesta:

WOZ: ciao

AN04: ciao

WOZ: in cosa posso esserti utile?

AN04: eh vorrei recuperare una vecchia prescrizione delle analisi che ho fatto
a giugno

WOZ: non ho capito, puoi ripetere?

AN04: vorrei recuperare, un a vecchia prescrizione delle analisi del sangue,
che ho fatto a giugno

Ancor meno frequente (6%) è la modificazione speculare all'estensione, cioè la cancellazione di elementi lessicali, probabilmente poiché percepiti come ridondanti e potenzialmente dannosi per la comprensione del sistema:

WOZ: ciao, in che cosa posso esserti utile?

AN14: <ee> vorrei sapere il saldo del mio conto corrente in banca

WOZ: non ho capito, puoi ripetere?

AN14: vorrei cap* conoscere il saldo del mio conto in banca

WOZ: ciao, ti serve qualcosa?

NV14: eh vorrei l'informazione sul credito<vv> residuo del mio cellulare* del
mio cellulare

WOZ: non ho capito, potresti ridirmelo?

NV14: vorrei sapere il credito del mio cellulare

6.4.5.3.2.2.2 Riformulazioni sintattiche La maggior parte delle riformulazioni (60%) presenta una variazione sintattica rispetto all'enunciato originario. Nel caso delle riformulazioni sintattiche, il parlante sembra individuare il problema non tanto a livello di comprensione dei singoli termini, ma a livello di presentazione della richiesta; questa può subire un'espansione, principalmente con lo scopo di esplicitare termini rimasti impliciti (o esplicitati durante l'interazione, ma che il parlante pensa il sistema possa non ricordare) o di aumentare il dettaglio della richiesta; oppure, può subire una riduzione, con l'eliminazione delle parti del discorso che possono essere

ridondanti o non rilevanti; o ancora può subire una sostituzione, nel caso sia la stessa formulazione della richiesta a poter essere problematica.

Il principale effetto delle espansioni sintattiche è l'esplicitazione di elementi del discorso, considerati acquisiti durante la conversazione o precedentemente ritenuti scontati relativamente alla richiesta. Nel caso gli elementi ritenuti acquisiti rimangano in ellissi, e il parlante si limiti a rispondere al sistema, una situazione d'errore può portare a riconsiderare la coesione dell'interazione e a esplicitare nuovamente il comando nella forma estesa:

WOZ: vuoi pagare subito il ticket con la carta di credito?

AN03: sì. quanto costa?

WOZ: non ho capito, puoi ripetere?

AN03: quanto costa pagare il ticket?

WOZ: sono rimaste ancora accese le luci della cucina e del salotto

NV09: spegni

WOZ: non riesco a comprendere, puoi ripetere?

NV09: dovresti spegnerle quelle della cucina e del salotto

Altrimenti, l'esplicitazione può riguardare parti del discorso ellittiche in quanto co-testuali, la cui ellissi è pragmaticamente accettabile nelle conversazioni quotidiane, e dove l'esplicitazione risulterebbe ridondante nella conversazione con un altro parlante:

WOZ: sono rimaste ancora accese le luci della cucina e del salotto

AN15: ah ok, allora dovresti spegnerle

WOZ: non riesco a comprendere, puoi ripetere?

AN15: allora dovresti spegnere le luci eh nelle stanze dove<vv> sono rimaste accese

WOZ: ecco, i film consigliati sono: i soliti ignoti, il sorpasso, l'armata
brancaleone, amici miei

AN13: vorrei sapere il costo

WOZ: non ho capito, puoi ripetere?

AN13: vorrei conoscere il costo di questi film

Tramite un'altra forma di espansione sintattica il parlante esplicita la parte del discorso legata alla stessa formulazione della richiesta, talvolta data per scontata durante l'interazione o formulata precedentemente dal sistema:

WOZ: sì, se vuoi puoi ordinarlo e pagarlo subito

NV05: sì <ee> con numero di carta

WOZ: non riesco a comprendere, puoi ripetere?

NV05: <ee> pagarlo, attraverso la carta

WOZ: vuoi una dieta dimagrante o una dieta depurante?

AN18: depurante

WOZ: non ho capito, # puoi ripetere? #

AN18: # vorrei # una dieta depurante

Altrimenti la forma piena non viene volutamente formulata dal parlante, fenomeno probabilmente legato al contesto d'interazione, essendo ogni enunciato dell'utente, di fatto, una richiesta di un'informazione o di un'azione da eseguire:

WOZ: va bene, aspetta un attimo va bene i termosifoni si accenderanno dalle
6 alle 22. ti serve altro?

NV13: la temperatura

WOZ: non ho capito, puoi ripetere?

NV13: voglio sapere la temperatura del riscaldamento

WOZ: va bene, la richiesta di fattura per entrambe le sedute è stata inviata

NV08: e il costo?

WOZ: non ho capito, puoi ripetere?

NV08: sì no volevo sapere quanto costava ogni seduta

Le espansioni sintattiche intrafrasali hanno un ruolo in parte diverso: coincidono con una maggiore formalità e una maggiore esplicitezza del discorso, in cui l'utente aggiunge elementiche porta l'utente ad aggiungere elementi per

specificare ciò di cui sta parlando. Gli enunciati così riformulati possono arrivare a essere ridondanti e potenzialmente problematici per un riconoscitore, concorrendo, più elementi, nel riempimento di un solo slot (cfr. 2.3.2), come nei primi due esempi:

WOZ: su rai 1 c'è il padrino che inizia alle 21. su canale 5 c'è via col vento
che inizia alle 21. su italia 1 c'è manhattan che inizia alle 20 e 45

AN12: bah via col vento alle <ee> ore ventuno

WOZ: non ho capito, potresti ridirmelo?

AN12: canale cinque, via col vento, ore ventuno

WOZ: su rai 1 c'è il padrino che inizia alle 21. su canale 5 c'è via col vento
che inizia alle 21. su italia 1 c'è manhattan che inizia alle 20 e 45.
posso fare qualcos'altro per te?

AN15: prenota la visione del padrino

WOZ: non ho capito, potresti ridirmelo?

AN15: prenota la visione del padrino su rai uno alle ore ventuno

WOZ: ciao, ti serve qualcosa?

NV03: sì dovrei impostare la temperatura del riscaldamento

WOZ: non ho capito, potresti ridirmelo?

NV03: dovrei impostare, la temperatura del riscaldamento di casa

Al confine tra l'aggiunta di ulteriori informazioni per precisare la richiesta e la specificazione di termini usati nel discorso possiamo trovare fenomeni di espansione dell'enunciato con proposizioni relative:

WOZ: va bene, la deviazione delle chiamate sul cellulare si attiverà a
mezzanotte

NV12: controlliamo il numero del cellulare?

WOZ: non ho capito, puoi ripetere?

NV12: sì vorrei controllare che il numero del cellulare che hai è esatto

WOZ: il fosamax70 serve a ridurre i rischi dell'osteoporosi, come fratture
vertebrali e dell'anca. ti serve altro?

AN22: sì, lo vorrei inserire nella lista dei farmaci

WOZ: non ho capito, puoi ripetere?

AN22: vorrei inserirlo, nella lista dei farmaci che prendo normalmente

Al contrario dell'espansione sintattica, da collocare all'interno di strategie di esplicitazione della coesione discorsiva, che si realizzano con la perifrasi di materiale ellittico, e, più in generale, di maggiore specificazione dei termini usati, la riduzione sintattica coincide il più delle volte con l'eliminazione delle ridondanze e di tutto ciò che a giudizio del parlante può essere problematico per la macchina in quanto già acquisito. Anche in questo caso, il primo meccanismo riguarda l'alleggerimento del carico informativo del singolo enunciato, per favorire l'esecuzione del compito:

WOZ: va bene, dovresti dirmi l'ora di accensione e i giorni della settimana

NV15: allora alle sette, dal lunedì al venerdì

WOZ: non ho capito, puoi ripetere?

NV15: alle sette di mattina

La strategia preferenziale di riduzione sintattica è l'ellissi del sintagma verbale, in favore della presentazione “asciutta” dei dati sensibili per lo svolgimento del compito; questo può accadere successivamente una domanda diretta del sistema (che rende ridondante la parte verbale):

WOZ: vuoi pagare ora con carta di credito o pagare alla consegna?

AN16: pagare con carta di credito

WOZ: non ho capito, puoi ripetere?

AN16: con la carta di credito

WOZ: ok, ho impostato l'assunzione dell'augmentin per tutta la settimana, alle 11 e alle 15. come vuoi essere avvisato?

AN07: voglio essere avvisata con un sms?

WOZ: non ho capito, potresti ridirmelo?

AN07: con un sms al cellulare

Oppure, successivamente a una domanda aperta, o più in generale quando il

parlante prende l'iniziativa; in tal caso, l'enunciato riformulato perde la parte verbale, focalizzando così le parti nominali sensibili. Infatti, è frequente che la riduzione avvenga successivamente a un elenco da cui l'utente deve selezionare degli elementi, in cui la reggenza verbale è implicita relativamente al contesto (che dà per scontata la selezione di una delle componenti dell'elenco):

WOZ: ecco i ristoranti con cucina romana aperti fino a tardi vicino casa tua: i
tre moschettieri, da zia maria, tutto alla brace

AN07: vorrei prenotare, stasera dopo le ventidue, per due persone da zia
maria

WOZ: non ho capito, puoi ripetere?

AN07: da zia maria, stasera dopo le ventidue, o alle ventidue, per due persone

WOZ: su rai 1 c'è il padrino che inizia alle 21. su canale 5 c'è via col vento
che inizia alle 21. su italia 1 c'è manhattan che inizia alle 20 e 45

NV04: <ee> fissami manhattan alle<vv> venti e quarantacinque

WOZ: non ho capito, potresti ridirmelo?

NV04: manhattan. alle venti e quarantacinque

In altri casi, differenti dall'elenco, l'eliminazione della testa verbale può essere ascritta più in generale a esigenze di alleggerimento e di semplificazione dell'enunciato. Dopo la domanda aperta, il problema per un sistema che deve essere in grado di gestire più domini e differenti azioni, può essere l'ambiguità dell'enunciato riformulato:

WOZ: ciao, in cosa posso esserti utile?

AN09: ho bisogno di una prescrizione per le analisi del sangue

WOZ: non ho capito, puoi ripetere?

AN09: prescrizione, analisi, sangue

AN13: ciao

WOZ: ciao, ti serve qualcosa?

AN13: vorrei sapere qual è il credito del mio cellulare

WOZ: non ho capito, potresti ridirmelo?

AN13: il credito del mio cellulare

Tra i meccanismi associabili alla riduzione sintattica abbiamo, inoltre, l'eliminazione di sintagmi e parti dell'enunciato marcatamente ridondanti:

WOZ: gli ultimi film di woody allen disponibili sono: incontrerai l'uomo dei tuoi sogni, midnight in paris, to rome with love

AN02: allora vorrei incontrerai l'uomo dei tuoi sogni, 4 euro, <ee> e to rome with love, 4 euro

WOZ: se ne acquisti 2 paghi 6 euro invece di 8

AN02: allora il primo e il terzo, cioè incontrerai l'uomo dei tuoi sogni, e to rome with love

WOZ: non ho capito, puoi ripetere?

AN02: incontrerai l'uomo dei tuoi sogni e to rome with love

WOZ: va bene, dovresti dirmi l'ora di accensione e i giorni della settimana

AN12: oh <ee> l'orario è alle otto. giorno da lunedì, alla domenica

WOZ: non ho capito, puoi ripetere?

AN12: dalle otto, dal lunedì, alla domenica

E la riduzione di costruzioni marcate in favore di costruzioni più lineari:

WOZ: ciao, ti serve qualcosa?

AN24: <ee> mi occorre <ee> sapere quanto ho di credito sul mio cellulare

WOZ: non ho capito, potresti ridirmelo?

AN24: <cc>mi serve, sapere, il credito del mio cellulare

Più in generale, alla specificazione ed esplicitazione presenti per le espansioni sintattiche, si contrappone un alleggerimento complessivo dell'enunciato riformulato, in cui possono venir cassati sintagmi di specificazione, meno rilevanti per il compito rispetto a quelli ritenuti sensibili per la macchina:

WOZ: roma - milano: la durata del viaggio è 3 ore e 45 minuti. milano - roma: la durata del viaggio è 3 ore e 35 minuti

AN11: vorrei acquistare il biglietto, di andata e ritorno
 WOZ: non riesco a comprendere, puoi ripetere?
 AN11: sì, vorrei acquistare il biglietto
 WOZ: va bene, per l'acquisto dovresti dirmi quale carta hai e il suo numero

WOZ: ciao, ti serve qualcosa?
 NV01: ho, prurito in tutto il corpo
 WOZ: non ho capito, potresti ridirmelo?
 NV01: ho, prurito
 WOZ: hai avuto gli stessi sintomi il 6 giugno 2012, il medico ti aveva
 prescritto l'halicar

A differenza delle forme di variazione viste prima, che possono aiutare a inquadrare la risoluzione degli errori entro due poli (banalmente, l'aver detto poco e l'aver detto troppo), la sostituzione di un sintagma con uno equivalente ha un inquadramento differente rispetto alla valutazione che fa il parlante riguardo le capacità di comprensione della macchina e a ciò che è problematico nell'enunciato. Con la sostituzione sintattica il parlante prova a cambiare la forma con cui presentare la richiesta, includendo (almeno parzialmente) una problematizzazione della cortesia e della distanza “sociale” dell'interagire con un macchina rispetto ai normali rapporti con un essere umano. Avviene infatti di frequente, tra le sostituzioni sintattiche, che il parlante cambi il modo della richiesta:

WOZ: va bene, aspetta un attimo. è spenta
 AN03: <ee> scusa mi puoi spegnere le le / la cucina e il salotto a mezzanotte?
 WOZ: non ho capito, puoi ripetere?
 AN03: sì, <ee> vorrei che tu spegnessi le luci del salotto e della cucina, a
 mezzanotte

WOZ: ciao, in cosa posso esserti utile?
 AN11: vorrei che mi leggessi le email
 WOZ: non ho capito, puoi ripetere?
 AN11: per cortesia se puoi leggermi le email arrivate

Contrariamente a quanto scritto in letteratura (ovvero, che i comandi dati al sistema, nelle occasioni di errore, tendono a essere più diretti rispetto agli originari) è difficile individuare una qualche tendenza generalizzata legata alla forza della richiesta, avviene infatti che l'enunciato riformulato abbia una maggiore forza illocutoria del precedente e minore cortesia:

WOZ: il medico ti ha prescritto l'assunzione del fosamax⁷⁰, per tre mesi una volta a settimana

AN13: perch*

WOZ: il fosamax⁷⁰ serve a ridurre i rischi dell'osteoporosi, come fratture vertebrali e dell'anca

AN13: potresti inserire, il farmaco nella mia lista di farmaci

WOZ: non ho capito, puoi ripetere?

AN13: inserisci il farmaco nella mia lista di farmaci

WOZ: va bene, ho disdetto l'appuntamento del 5 dicembre alle 14

AN16: puoi prenderne un altro per il sette dicembre, alle quindici

WOZ: non ho capito, potresti ridirmelo?

AN16: ri* prendi un altro appuntamento col dentista, per il sette dicembre, alle ore, quindici

WOZ: la prossima puntata di che tempo che fa è sabato alle 21

AN18: me la puoi registrare?

WOZ: non ho capito, potresti ridirmelo?

AN18: eh registra la puntata di che tempo che fa

WOZ: la puntata di che tempo che fa di sabato verrà registrata

Contrariamente a quanto ci si aspetterebbe (e a quanto detto riguardo alla non rilevanza della funzione di cortesia, cfr. Johnstone et al. 1994), sono presenti anche casi in cui l'enunciato riformulato è meno diretto dell'originario⁷⁰:

⁷⁰ Non è questo il luogo, ma la riflessione sulla cortesia nell'interazione uomo-macchina non può non tenere conto delle abitudini culturali del parlante: è molto raro, per un parlante, usare l'imperativo nella propria quotidianità sociale; così, nella quotidianità avviene che una richiesta troppo "sgarbata" (per fretta o noncuranza) a fronte della perplessità del nostro interlocutore, sia rimodulata in maniera più cortese.

WOZ: sono rimaste ancora accese le luci della cucina e del salotto

NV01: spegnile allora

WOZ: non riesco a comprendere, puoi ripetere?

NV01: se le puoi spegnere per favore

WOZ: ok, tutte le luci di casa sono spente

WOZ: ecco i ristoranti con cucina romana aperti fino a tardi vicino casa tua: i
tre moschettieri, da zia maria, tutto alla brace

NV15: prenota da zia maria, per due persone, alle venti di stasera

WOZ: non ho capito, puoi ripetere?

NV15: puoi prenotare da zia maria?

WOZ: vuoi sapere com'è regolato il riscaldamento?

NV07: sì

WOZ: è regolato a 22 gradi

NV07: cambia temperatura

WOZ: non ho capito, puoi ripetere?

NV07: voglio cambiare la temperatura

Ragionare sul modo con cui porre una richiesta porta con sé un altro aspetto che il parlante considera rilevante nell'interazione, ovvero l'allocuzione inclusa nei tipi di richiesta: a chi mi devo rivolgere? Il cambio di allocuzione tra l'enunciato originario e quello riformulato, presente nel 21% (43 casi) delle sostituzioni sintattiche. Più della metà dei cambi di allocuzione portano il parlante a formulare la frase in prima persona, come desiderio o necessità, a partire da una domanda precedentemente rivolta al sistema dandogli del *tu*:

WOZ: ecco i ristoranti con cucina romana aperti fino a tardi vicino casa tua: i
tre moschettieri, da zia maria, tutto alla brace

AN13: bene, alle<vv> ventidue<vv> puoi prenotare <ee> un tavolo per due
persone?

WOZ: non ho capito, puoi ripetere?

AN13: vorrei prenotare un tavolo per due persone alle, 22

WOZ: va bene, in quale ristorante?

In un solo caso il parlante riformula in prima persona una richiesta formulata in 3° persona (raramente usata dai parlanti per rivolgersi al sistema):

WOZ: roma - milano: la durata del viaggio è 3 ore e 45 minuti

AN16: può prenotare il biglietto a nome di, maria rossi

WOZ: non riesco a comprendere, puoi ripetere?

AN16: vorrei prenotare il biglietto, a nome di maria rossi

In un altro singolo caso viene sostituita la 4° persona, altro fenomeno raro nel *corpus*:

NV12: controlliamo il numero del cellulare?

WOZ: non ho capito, puoi ripetere?

NV12: sì vorrei controllare che il numero del cellulare che hai è esatto

Altrimenti l'originaria prima persona può essere sostituita da una richiesta impersonale, mentre meno frequente è il passaggio dall'allocuzione a una richiesta impersonale (come nel secondo esempio):

WOZ: roma - milano: la durata del viaggio è 3 ore e 45 minuti

AN02: e per la prenotazione del biglietto?

WOZ: non riesco a comprendere, puoi ripetere?

AN02: dovrei prenotare il biglietto

WOZ: va bene, per la prenotazione dovresti darmi il tuo nominativo e un numero di cellulare

WOZ: bene, la tua dieta verrà elaborata entro 48 ore. ti serve altro?

AN08: inserisci nel profilo personale

WOZ: non ho capito, puoi ripetere?

AN08: inserire nel profilo personale

In un solo caso l'allocuzione viene modificata, dalla prima persona originaria, con

una richiesta in seconda persona:

WOZ: l'antifurto si attiverà domani alle 11. ti serve altro?

AN22: sì, dovrei deviare le telefonate del telefono di casa, da questa sera, a mezzanotte, sul mio cellulare

WOZ: non ho capito, puoi ripetere?

AN22: dovrei deviare, le telefonate del telefono di casa, sul mio cellulare, da stasera, a mezzanotte

WOZ: va bene, vuoi impostarla sul tuo cellulare o su un altro numero?

6.4.5.3.2.2.3 Questioni testuali e pragmatiche. Alle variazioni sintattiche e lessicali a cui fa affidamento il parlante per andare incontro al sistema, corrisponde spesso, come abbiamo visto, una problematizzazione della coesione interna all'interazione. Sul piano testuale, infatti, sono da ricercarsi le ragioni di molte delle riformulazioni sintattiche: all'espansione corrisponde spesso l'esplicitazione di strutture sottintese nell'enunciato precedente alla riformulazione, sia perché dette dal sistema, sia perché implicite nel contesto del dialogo; allo stesso modo, può avvenire, sebbene più raramente, il processo inverso, ovvero che nell'enunciato riformulato sia assente un costrutto considerato acquisito e la cui ridondanza può essere giudicata problematica per la comprensione da parte del sistema.

Abbiamo visto, relativamente alla sostituzione lessicale, la frequenza del ripristino del referente di un pronome anaforico durante una situazione di mancata acquisizione. A differenza delle ellissi, la cui parafrasi oscilla tra ripristino ed espansione, ed è quindi difficilmente quantificabile all'interno del *corpus*, per i pronomi anaforici possiamo valutare da un punto di vista oggettivo la portata della tendenza ad esplicitare e recuperare parti del discorso date per implicite al momento della formulazione dell'input. Nel caso l'enunciato originario presenti un pronome anaforico (65 casi), nel 74% dei casi il pronome viene sostituito con l'antecedente lessicale:

WOZ: sono rimaste ancora accese le luci della cucina e del salotto

AN03: dobbiamo spegnerle

WOZ: non riesco a comprendere, puoi ripetere?

AN03: do* dobbiamo spegnere le luci in cucina e in salotto

WOZ: il farmaco costa 4,50 euro
NV04: puoi prenotarmelo?
WOZ: non riesco a comprendere, puoi ripetere?
NV04: puoi prenotare il farmaco?

WOZ: su rai 1 c'è il padrino che inizia alle 21. su canale 5 c'è via col vento
che inizia alle 21. su italia 1 c'è manhattan che inizia alle 20 e 45
NV12: mi<vv> puoi programmare per favore quello sul primo?
WOZ: non ho capito, potresti ridirmelo?
NV12: sì, puoi per favore, <mm> programarmi il film su<vv> sul primo?

Nel restante minor numero di casi in cui è presente un pronome anaforico nell'enunciato non compreso dal sistema, il pronome viene riproposto anche nell'enunciato riformulato. In tutto il *corpus* è presente solo un caso in cui il pronome viene immesso successivamente alla riformulazione e il cui referente è presente nell'enunciato non compreso dal sistema: in tal caso, la domanda del sistema, che suggerisce entrambe le opzioni relativamente all'acquisto o alla prenotazione del biglietto, può aver indotto il parlante a focalizzare la propria attenzione sulla parte rilevante del proprio enunciato, piuttosto che su una parte ritenuta acquisita (e potenzialmente problematica in quanto ridondante):

WOZ: roma - milano: il costo del biglietto di seconda classe è 45 euro
milano- roma: il costo del biglietto di seconda classe è 48 euro se vuoi
puoi procedere con l'acquisto del biglietto tramite carta di credito
altrimenti puoi prenotare il biglietto
AN03: voglio prenotare il biglietto
WOZ: non riesco a comprendere, puoi ripetere?
AN03: devo prenotarlo

6.4.6. Situazioni di errata acquisizione

Oltre agli errori dovuti alla non comprensione, sono state simulate le situazioni in cui il sistema comprende un altro input rispetto a quello detto dal

parlante. Se il sistema individua subito un potenziale errore, può chiedere conferma dell'acquisizione, in caso contrario, lo svolgimento dell'attività può essere viziata, con risultati differenti da quelli prefissati dell'utente. La maggior parte degli output di errata acquisizione sono richieste di conferma esplicita, la strategia scelta per il controllo delle acquisizioni del sistema (rispetto ad altre meno robuste, come descritto in 6.4.3). Le richieste di conferma esplicita presentano almeno una singola acquisizione errata (che può riguardare uno dei dati da fornire al sistema o la stessa azione da svolgere); inoltre, pur non rientrando nel dominio del controllo di possibili errori, nella simulazione sono presenti anche due feedback di completamento con un'errata acquisizione.

Ogni richiesta di conferma è stata preparata in relazione al compito e ai dati che il soggetto doveva fornire per il suo completamento (cfr. capitolo 4); quindi, il loro uso è stato in parte prefissato, in quanto dipendente dal compito che il soggetto doveva svolgere, in parte affidato alle scelte del wizard, a seconda della complementarità con il turno precedente. Ad esempio, se nel compito era prevista la prenotazione di un biglietto ferroviario per una determinata destinazione, il wizard poteva usare la richiesta di conferma per una destinazione diversa da quella prevista dal compito (e quindi errata) all'interno del sottocompito di ricerca dei treni disponibili.

6.4.6.1 Sequenza di riparazione

A differenza delle richieste di ripetizione, che comportano una sequenza di risoluzione in due turni, la sequenza di risoluzione che si presenta nel caso di richieste esplicite è variabile, a seconda del turno successivo all'errore. Questo, infatti, può essere risolto in due turni, nel caso l'utente, oltre a disconfermare l'acquisizione del sistema, proponga la correzione nello stesso turno, senza aspettare che il sistema prenda nuovamente il turno per richiedere l'informazione rifiutata:

WOZ: di quanti euro vuoi ricaricarlo?

AN22: di venti euro

WOZ: vuoi fare una ricarica di cento euro?

AN22: no di venti euro

WOZ: ciao come posso aiutarti?

NV12: ho bisogno di un farmaco per l'emicrania

WOZ: vuoi un farmaco per il mal di schiena?

NV12: no, per il mal di testa

Altrimenti l'errore può essere risolto in quattro turni se il soggetto, nel turno successivo alla richiesta di conferma, si limita a segnalare l'errata acquisizione e ad aspettare una nuova richiesta da parte del sistema per poter fornire l'informazione corretta:

WOZ: di quanti euro vuoi ricaricarlo?

NV06: venti

WOZ: vuoi fare una ricarica di 100 euro?

NV06: no

WOZ: di quanti euro vuoi ricaricarlo?

NV06: venti euro

WOZ: ciao, ti serve qualcosa?

AN12: sì vorrei una lista del, del film dei de woody allen

WOZ: vuoi cercare tra gli ultimi film di martin scorsese?

AN12: no

WOZ: in base a cosa devo cercare?

AN12: woody allen

A seguito di una richiesta di conferma esplicita la strategia preferenziale è la correzione nel turno successivo (94%)⁷¹. Va notato, tuttavia, che la maggior parte delle risoluzioni in quattro turni avviene successivamente a una richiesta di conferma di un numero di carta di credito (erroneamente acquisito dal sistema): questo particolare caso mostra come la risoluzione di un errore possa variare a seconda dalla rilevanza che viene data all'attività. Infatti, durante la conferma del numero di carta di credito (errore introdotto in un compito sul pagamento di un ticket), aumenta la percentuale dei casi in cui il parlante attende istruzioni dal sistema invece di

⁷¹ Differenziandosi in maniera speculare al corpus C-ORAL-ROM, dove è ampiamente preferita la semplice disconferma in attesa di istruzioni del sistema.

correggere direttamente l'errore: se nel resto del *corpus* solo il 6% di richieste di conferma esplicita si risolvono su iniziativa del sistema, nel caso dell'errore durante la transazione la percentuale sale al 37%:

WOZ: qual è il numero della carta di credito?

NV16: quattro nove . sette sette . cinquanta , cinquantasei . diciotto ventidue

WOZ: lo ripeto . quarantanove settantasette cinquanta cinquantasei
diciannoveventiquattro , è corretto?

NV16: no

WOZ: qual è il numero della carta di credito?

NV16: quattro nove . sette sette . cinquanta . cinquantasei . diciotto ventidue

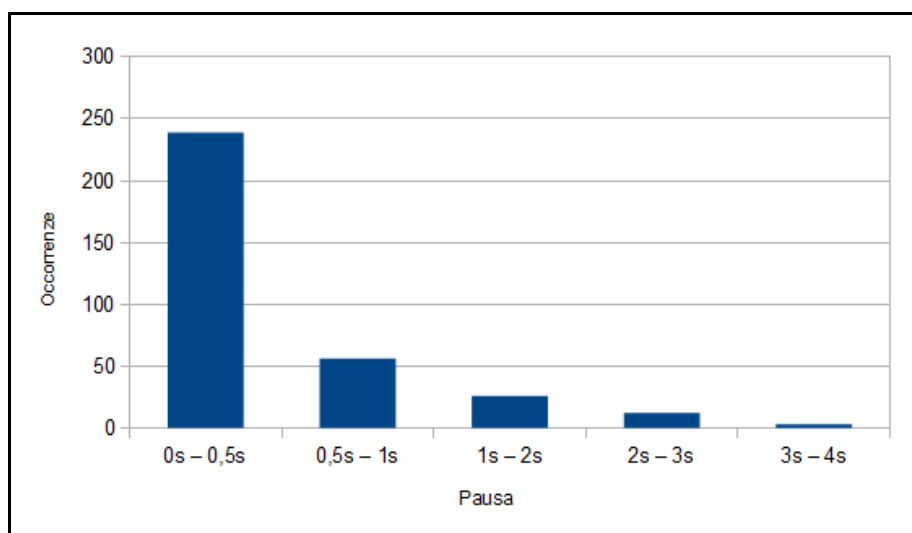
WOZ: lo ripeto , quarantanove settantasette cinquanta cinquantasei diciotto
ventidue , è corretto?

NV16: sì

WOZ: bene , il ticket è stato pagato

Naturalmente, l'iniziativa nella risoluzione dell'errore non dipende solo dall'atteggiamento del parlante, ma anche dal sistema e dalla gestione di presa del turno: quando corregge nello stesso turno, l'utente può aver atteso comunque una possibile richiesta del sistema dopo la sua disconferma, provvedendo alla correzione solo successivamente alla mancata presa del turno da parte del sistema; al contrario, una presa del turno molto veloce seguentemente alla disconferma del parlante, obbliga quest'ultimo a dover seguire l'iniziativa del sistema.

Nella simulazione, per non avere casi in cui l'iniziativa del sistema è obbligata da una veloce presa del turno, abbiamo definito un tempo di attesa intorno ai 3 secondi dopo la disconferma, prima di prendere il turno e richiedere nuovamente l'informazione errata. Si nota così, nel corpus, una notevole presenza di pause che intercorrono tra la disconferma e la correzione, di cui il 30% sono superiori al mezzo secondo. In media, le pause successive alla disconferma sono di 0,5 secondi: la lunghezza della pausa è direttamente proporzionale alla probabilità che il parlante stia attendendo una nuova richiesta e che sia quindi meno propenso a prendere il turno. Da un punto di vista applicativo, questo porta a dover considerare la pausa successivamente alla disconferma per poter tarare una diversa tempistica di presa del turno così da evitare partenze simultanee.



Graf. 26: Durata della pausa tra disconferma e correzione

Anche durante la risoluzione di un'errata acquisizione il parlante può sovrapporsi al sistema: la possibilità è strettamente connessa alla posizione dell'errore durante il turno del sistema. Delle richieste di conferma esplicita la maggior parte presenta l'informazione errata da confermare nella posizione finale del turno: in questi casi non si riscontra alcuna sovrapposizione; tuttavia, anche quando l'errore non è situato nella parte finale del turno⁷² (39% delle richieste di conferma usate), i parlanti mostrano l'usuale "pazienza" nell'aspettare la fine del turno del sistema anche durante una richiesta di conferma esplicita, eccetto 5 casi su 196 in cui il parlante interviene prima che sia finito il turno del sistema:

AN24: sì. mi vai sulla guida dei film che eh hh faranno questa sera in
televisione, perché ho deciso di non uscire

WOZ: vuoi vedere la programmazione sportiva per # stasera in tv #?

AN24: # <nn>no # la i film in
programmazione dopo le ore venti

WOZ: va bene, puoi dirmi i giorni?

AN15: a giorni alterni

WOZ: vuoi che le piante vengano irrigate una volta a settimana, dalle 23 alle
23 e # 10 #?

⁷² Sono state incluse perciò anche le richieste di conferma esplicita con più di un dato errato, poiché il primo errore è da considerarsi intermedio e non finale.

AN15: # no, # vor* a giorni alterni

WOZ: va bene, ho impostato l'irrigazione delle piante a giorni alterni, dalle
23 alle 23 e 10

Anche in questo caso, le richieste di conferma del numero di carta di credito comportano una situazione diversa: sono infatti contraddistinte da una breve pausa (0,4s) tra la ripetizione del numero di carta di credito (errato) e la richiesta di conferma esplicita, che favorisce l'intervento prima della richiesta. Il parlante prova così a prendere il turno subito dopo l'errore del sistema, prima ancora che questi ne faccia richiesta esplicita, in più della metà dei casi (15 su 27):

WOZ: # qual # è il numero della carta di credito?

AN12: quattro nove (0.6) sette sette (0.5) cinque zeo cinque sei (0.6) uno otto
due due

WOZ: (2.8) lo ripeto (0.4) quarantanove settantasette cinquanta cinquantasei
diciannove # ventiquattro #

AN12: # no # no # è sbagliato #

WOZ: # è corretto? #

AN12: no è sbagliato

6.4.6.2 Individuazione dell'errore

A differenza delle richieste di ripetizione, che indicano inequivocabilmente un errore o una mancata ricezione da parte del sistema, le richieste di conferma esplicite possono, ovviamente, presentare un'acquisizione corretta. Nel *corpus* sono presenti casi in cui il parlante conferma un'acquisizione differente da quella detta precedentemente: per quanto tale situazione si possa verificare anche nell'interazione con un sistema reale in un contesto reale, vanno ricordati, in questo caso, i vizi dell'ambiente sperimentale, che comporta generalmente delle esigenze e un'attenzione minori per via dell'assenza di interessi reali per il successo dell'interazione. Le conferme di un'acquisizione sbagliata ricoprono l'8% delle richieste di conferma:

WOZ: ciao, come posso aiutarti?

AN13: vorrei prenotare un treno per milano

WOZ: vuoi un treno per merano?

AN13: sì

AN26: bene, allora pagheremo con la mia carta di credito , vuoi il numero?

WOZ: sì

AN26: quattro nove sette sette . cinquanta cinquantasei , diciotto ventidue

WOZ: lo ripeto: quarantanove settantasette cinquanta cinquantasei

diciannove ventiquattro

AN26: perfetto

WOZ: è corretto?

AN26: correttissimo, brava

WOZ: bene, il ticket è stato pagato

6.4.6.3 Strategie di correzione a confronto

Dopo una richiesta di conferma esplicita, il parlante può disconfermare l'informazione acquisita e correggere il sistema. Abbiamo già visto i casi in cui viene solo disconfermata l'acquisizione e viene attesa la nuova richiesta per correggere l'errore: la correzione in questi casi è distribuita sui due turni del parlante, che corrispondono ad altrettanti atti linguistici. Quando invece la correzione avviene nel turno successivo, sono presenti quattro configurazioni del turno diverse.

6.4.6.3.1 Disconferma + correzione

Il turno successivo a un'errata comprensione del sistema presenta, di preferenza, un feedback di disconferma a cui segue la correzione dell'acquisizione (spesso inframezzata da una lunga pausa, cfr. paragrafo 6.1). Il feedback è quasi sempre *no*, talvolta ripetuto (la ripetizione del *no* può essere anche essere considerata una peculiarità di alcuni parlanti, ricorrendo solo in due soggetti), e talvolta *sbagliato* (nel caso la domanda di conferma sia *è corretto?*). La disconferma seguita dalla correzione copre il 72% dei casi:

WOZ: ciao, in che cosa posso esserti utile?

NV13: un mal di testa

WOZ: vuoi un farmaco per il mal di schiena?

NV13: no per il mal di testa

WOZ: va bene, per l'acquisto dovresti dirmi quale carta hai e il suo numero

AN: la carta di credito è visa , il numero è quarantanove sette sette ,
cinquanta, cinquantasei , diciotto ventidue

WOZ: lo ripeto: quarantanove settantasette cinquanta cinquantasei
diciannove ventiquattro

AN23: no

WOZ: è corretto?

AN23: no, è sbagliato <ee> diciotto ventidue

In due casi, il feedback di disconferma è anticipato da un'affermazione, che svolge contemporaneamente le funzioni di feedback d'accordo e di presa del turno:

WOZ: ciao, in che cosa posso esserti utile?

AN13: allora avrei bisogno di impostare la radiosveglia tutte le mattina, alle
sette, da lunedì al venerdì. e impostarla anche su un canale<ee> radio
dove trasmettono notizie

WOZ: vuoi che si accenda alle 7, ogni giorno della settimana?

AN13: sì, no dal lunedì al venerdì

6.4.6.3.2 Disconferma esplicita + correzione

In molti meno casi (4%), l'errore non viene marcato dal solo feedback di disconferma, ma viene esplicitamente ripresa l'informazione errata, a cui di norma segue la correzione:

WOZ: ciao

AN12: ciao # buongiorno #

WOZ: # ti serve qualcosa? #

AN12: eh sì <ee> avrei voglia di vedere stasera qualche bel film, mi puoi far
vedere la lista per favore

WOZ: vuoi vedere la programmazione sportiva per stasera in tv?

AN12: nono, sportiva no, film

Solo in due casi la correzione anticipa la disconferma esplicita dell'errata acquisizione:

WOZ: ciao, come posso aiutarti?

AN17: avrei bisogno di un farmaco, per l'emicrania

WOZ: vuoi un farmaco per il mal di schiena?

AN17: per l'emicrania, non mal di schiena

6.4.6.3.3 *Correzione*

La seconda strategia più usata (18%) è la diretta correzione dell'acquisizione errata, senza che questa sia marcata da alcun feedback di disconferma:

WOZ: ciao, come posso aiutarti?

AN12: ho un forte mal di testa, che farmaco dovrei pre* potrei prendere?

WOZ: vuoi un farmaco per il mal di schiena?

AN12: per la testa

WOZ: non ho capito, puoi ripetere?

NV03: la programmazione del<cc>, dell'irrigazione, del terrazzo deve essere aperto alle ventitre e chiuso alle ventire e dieci, di un giorno sì e un giorno no

WOZ: vuoi che le piante vengano irrigate una volta a settimana, dalle 23 alle 23 e 10?

NV03: <ee> un giorno sì e un giorno no, dalle ventitre alle ventitre e dieci

6.4.6.3.4 *Precisazione*

Oltre ai casi già visti, in cui con la disconferma (implicita o esplicita) si marca l'individuazione dell'acquisizione come errore (ovvero, in cui si corregge un output non previsto rispetto a quanto detto), sono presenti altri in cui l'individuazione non è

altrettanto chiara.

In tali casi, il parlante non disconferma l'acquisizione del sistema, ma la completa, ripetendo (o introducendo) l'informazione corretta: l'output del sistema non viene quindi giudicato errato, ma incompleto. Tale situazione si verifica in particolare nel caso di una richiesta di conferma presente in due compiti nel dominio domotico (controllo delle luci di tutta la casa e disattivazione dell'impianto dell'acqua), in cui il sistema presenta un'acquisizione sbagliata (ovvero non comprende “tutta la casa”, ma solo una delle stanze di casa), e il parlante, invece di disconfermare, ripete la precedente richiesta. Nell'esempio si può vedere come tali situazioni possano essere strutturalmente ambigue, in quanto l'ingresso fa parte dell'appartamento; quindi il controllo delle luci in tutta casa comporta anche il controllo delle luci dell'ingresso:

WOZ: ciao, come posso aiutarti?

AN07: devi controllare le luci <ii> nella casa se sono rimaste accese

WOZ: vuoi controllare se le luci dell'ingresso sono accese?

AN07: sì in tutta la casa

Nel seguente esempio la specificazione è particolarmente evidente: il parlante non corregge il sistema, ma esplicita la propria richiesta scomponendola (*anche in cucina*) per rendere meno ambiguo il comando, poi ripetuto:

WOZ: non ho capito, puoi ripetere?

AN07: sì dovresti chiudere l'acqua, in tut* per tutto l'appartamento / in tutta la casa

WOZ: vuoi chiudere l'impianto in bagno?

AN07: sì. anche in cucina, in tutto l'appartamento in tutta la casa

6.4.6.4 Proprietà della correzione

6.4.6.4.1 Riprogrammazioni del compito

Quando il sistema chiede esplicitamente conferma di un'acquisizione, lo

svolgimento dell'attività si interrompe mentre l'interazione prosegue per dare modo all'utente di confermare o disconfermare quanto compreso nel turno successivo, deputato di preferenza allo scioglimento delle possibili incomprensioni.⁷³ Tuttavia, pur se in misura minore rispetto alle situazioni di non comprensione da parte del sistema (cfr. 5.3.2.1), anche nel caso di richiesta conferma esplicita può esserci un carico informativo diverso rispetto ai turni precedenti del parlante. Infatti, quando il parlante, con la disconferma, presenta una nuova informazione rispetto al suo turno precedente, il più delle volte è stimolato da un errore del sistema, il cui output presenta un'acquisizione, non detta dal parlante, relativa a un altro *slot*⁷⁴: questo porta il parlante a riprogrammare la risoluzione del compito correggendo il sistema relativamente a quella variabile dato, come avviene nel 9% dei casi:

WOZ: ciao, in che cosa posso esserti utile?

NV08: vorrei andare al ristorante

WOZ: cerchi un ristorante cinese?

NV08: no cucina romana

WOZ: ecco i canali che trasmettono musica classica: filodiffusione, radio classica, radio tre

NV15: radio tre

WOZ: va bene, ho impostato l'impianto stereo su radio tre vuoi si spenga tra dieci minuti?

NV15: alle ventitre

È invece più raro (4 casi) che venga semplicemente disconfermata l'errata acquisizione, senza che venga corretta la nuova informazione, o perché ritenuta opzionale e non rilevante, o in attesa di una nuova domanda del sistema:

WOZ: ciao, in che cosa posso esserti utile?

AN06: <ee> una lista di ristoranti vicino a casa aperti fino a tardi

WOZ: cerchi un ristorante cinese?

⁷³ Nelle varie classificazioni di atti di dialogo elaborate nell'ambito del dialogo uomo-macchina questi sono i turni che generalmente prevedono un'azione "indietro" (*back-looking acts*, cfr. 2.4.5.1) rispetto allo svolgimento dell'attività.

⁷⁴ In tale caso, l'errore avviene per via del riconoscitore vocale, che, rispetto a quanto detto, individua un'altra parola, rilevante per il completamento di uno degli slot attivi di comprensione.

AN06: no

Altrettanto rari (5 occorrenze) sono i casi in cui l'utente presenta nuove informazioni deliberatamente durante un'errata comprensione:

WOZ: ciao, in che cosa posso esserti utile?

NV02: puoi spegnere, luci e gas, a* a casa? per domani?

WOZ: vuoi programmare l'accensione della caldaia?

NV0: no, spegnere le luci di casa. entro le ore undici, a casa

Come ci si può aspettare dalla richiesta esplicita, che ancora la conversazione al problema occorso, è raro che l'utente presenti nuove informazioni o cambi la programmazione della risoluzione del compito. Esclusi i casi in cui il parlante accetta la richiesta di conferma errata (motivabili semplicemente con la differente attenzione che si può avere durante lo svolgimento di un compito), ci sono solo 5 casi in cui il parlante cambia progetto rispetto ai turni precedenti:

NV16: imposta la temperatura per la stanza da letto

WOZ: ok, che temperatura vuoi?

NV16: ventitré gradi, anzi no. venticinque

WOZ: vuoi impostare la temperatura del salotto a venticinque gradi?

NV16: no, della stanza da letto, a ventiquattro

Non ha invece senso parlare di riprogrammazioni pragmatiche con un carico informativo minore: essendo il turno di correzione pertinente unicamente per alcune informazioni dette in precedenza, e costituendo di fatto un sottocompito, la correzione presenta quasi sempre meno informazioni rispetto a quelle dette in precedenza, essendo focalizzata principalmente sull'informazione errata.

6.4.6.4.2 Formulazione intera o parziale

A fronte della richiesta di conferma del sistema, si riscontrano due macrotipi di correzione: il parlante può ripresentare la richiesta completa o altrimenti correggere unicamente le parti dell'enunciato erroneamente acquisite. Questa seconda

strategia è preferenziale, ricoprendo più della metà del *corpus* (71%) casi. Nella maggior parte dei casi (il 56% delle correzioni della sola acquisizione errata) la correzione è costruita con parte dello stesso sintagma della richiesta di conferma:

WOZ: di quanti euro vuoi ricaricarlo?

AN22: venti euro

WOZ: vuoi fare una ricarica di cento euro?

AN22: no di venti euro

WOZ: ciao, come posso aiutarti?

AN16: avrei bisogno, di un'informazione sul treno, da roma a milano

WOZ: vuoi un treno per merano?

AN16: per milano

Con un frequenza leggermente minore (44%), la correzione del parlante viene presentata con il resto del sintagma eliso:

AN24: avrei bisogno di sapere gli orari dei treni per milano, per il giorno
ventisette aprile al mattino, e <vv>il ritornare il giorno dopo sempre in
mattinata il ventotto

WOZ: vuoi un treno per merano?

AN24: no milano

WOZ: ecco l'e-mail. da: fabio bianchi. ciao, senti, per stasera è confermata la
cena per le ventuno a casa mia, solo un problema: non posso passare a
prenderti

AN16: mi dovresti prenotare un taxi, alle ore venti e trenta

WOZ: ti serve un taxi per le diciassette?

AN16: no venti e trenta

La riproposizione di un enunciato completo a seguito della richiesta di conferma avviene invece più raramente (29%). In questi casi, come vedremo più avanti, spesso la richiesta ricalca quella del sistema:

WOZ: ciao, come posso aiutarti?
AN13: ho un gran mal di testa
WOZ: vuoi un farmaco per il mal di schiena?
AN13: no ho / voglio un farmaco per il mal di testa

WOZ: per quando ti serve?
AN21: <ee> per le venti e trenta
WOZ: ti serve un taxi per le 17?
AN21: no, mi serve un taxi per le venti e trenta

In particolare, la richiesta completa viene riformulata quando l'errore è legato alla testa verbale della richiesta di conferma, e riguarda non tanto una singola informazione ma l'azione da svolgere:

WOZ: ciao, come posso aiutarti?
AN18: ho bisogno dell'elenco dei film registrati dalla tv
WOZ: vuoi registrare un film?
AN18: no voglio vedere l'elenco dei film registrati ieri

WOZ: ciao, in che cosa posso esserti utile?
AN13: devo acquistare un farmaco che mi ha prescritto il medico, e vorrei sapere, il nome, e a cosa serve
WOZ: vuoi sapere quando puoi andare dal medico?
AN13: no, voglio sapere il nome di un farmaco che mi ha prescritto il medico, e voglio sapere a cosa serve

6.4.6.4.3 Fenomeni di adattamento

A differenza delle richieste di ripetizione, con cui il sistema non dà alcuna informazione al parlante sulle cause dell'errore, l'errata acquisizione, oltre a presentare esplicitamente l'informazione che necessita una conferma o una disconferma, fornisce un modello (lessicale e sintattico) a cui il parlante può o meno adattarsi per cercare di farsi comprendere dal sistema. I fenomeni di adattamento del lessico e della sintassi al proprio interlocutore sono già stati ampiamente documentati

in letteratura: si veda a tal proposito Gustafsson et al. 1997, Parent – Eskenazi 2010 (dal punto di vista lessicale), Branigan et al. 2003 e Branigan et al. 2006 (dal punto di vista sintattico), Stoyanchev – Stent 2009 (dal punto di vista lessicale e sintattico). Si veda inoltre Porzel – Baudis 2004, per i quali, proprio per la presenza di fenomeni di adattamento, bisogna cominciare a prestare maggiore attenzione per il modulo di generazione del linguaggio, su cui finora la ricerca (da un punto di vista progettuale) è stata abbastanza carente.

Dopo la richiesta di conferma, il parlante può adattarsi alla forma della richiesta del sistema, al lessico, ripetere il sintagma o il lessico usato o riformulare senza alcun adattamento. Rispetto alle richieste di ripetizione, che comportano un bassa presenza di ripetizioni in favore di riformulazioni operate sulla base delle presupposizioni del parlante, nel caso delle richieste di conferma esplicita la strategia più frequente è proprio la ripetizione, che spesso trova accordo lessicale e sintattico con la richiesta di conferma del sistema:

WOZ: va bene, la temperatura interna è venti gradi, quella esterna dieci gradi

AN15: ok, è possibile impostare la temperatura in camera da letto, a ventitré gradi?

WOZ: vuoi impostare la temperatura del salotto a venticinque gradi?

AN: no, in camera da letto, a ventitré gradi

WOZ: ok, quando devi prendere l'augmentin?

AN16: per tutta la settimana

WOZ: va bene, puoi indicarmi anche l'orario?

AN16: alle undici e alle quindici

WOZ: vuoi fissare l'augmentin per lunedì, mercoledì e venerdì alle 9?

AN16: no. alle undici e alle quindici, per tutta la settimana

In 106 casi è presente almeno un adattamento alla richiesta di conferma del sistema (dove per adattamento intendiamo naturalmente i casi in cui la forma o il lessico con cui era stata presentata la richiesta si differenziavano dalla forma e dal lessico usato dal sistema). I casi di adattamento sintattico li abbiamo già parzialmente inquadrati a proposito delle richieste complete rivolte al sistema, o di parte del sintagma ripreso nella correzione; lo stesso può avvenire a livello lessicale, quando il

parlante cambia i termini usati, presupponendo che il sistema comprenda il lessico che usa:

WOZ: ciao, come posso aiutarti?

AN15: dovresti <ee> chiudere le tapparelle, in salotto

WOZ: vuoi abbassare le tapparelle della cucina?

AN15: <ee> dovresti, abbassare, le tapparelle, in salotto

WOZ: ciao come posso aiutarti?

NV12: ho bisogno di un farmaco per l'emicrania

WOZ: vuoi un farmaco per il mal di schiena?

NV12: no, per il mal di testa

Mentre è più raro che la riformulazione da parte del parlante non si appoggi alla richiesta del sistema:

WOZ: da che ora a che ora deve essere acceso?

AN15: alle otto di mattina

WOZ: vuoi che si accenda alle 16 del pomeriggio?

AN15: <cc>no <ee> l'ora di accensione è alle otto di mattina

WOZ: la temperatura interna è 20 gradi

NV16: puoi impostare in camera da letto, ventitré gradi

WOZ: vuoi impostare la temperatura del salotto a 25 gradi?

NV16: no. stanza da letto

La tendenza all'adattamento da parte dell'utente alla forma con cui il sistema presenta la propria richiesta è ancora più evidente nella dizione dei numeri. Nel solito caso di conferma del numero di carta di credito, avviene di frequente che il parlante adatti la dizione dei numeri della carta a quella del sistema (per decimali). Dei 28 casi (da cui sono stati sottratti i 7 con la medesima dizione) in cui il parlante pronuncia il numero di carta diversamente dal sistema, in 20 casi adatta la propria dizione a quella per decimali del sistema:

AN17: per il pagamento . farei con la carta visa . numero , quattro nove sette
sette , cinque zero , cinque sei , uno otto due due

WOZ: lo ripeto , quarantanove settantasette cinquanta cinquantasei
diciannove ventiquattro

AN17: no è # sba* #

WOZ: # è # corretto?

AN17: è sbagliato . ripeto

WOZ: qual è il numero della carta di credito?

AN17: quarantanove . settantasette , cinquanta , cinquantasei , diciotto ,
ventidue

WOZ: qual è il numero della carta di credito?

AN04: quattro nove , sette sette , cinque zero cinque sei , uno otto due due

WOZ: lo ripeto , quarantanove settantasette cinquanta cinquantasei
diciannove # ventiquattro #

AN04: # no # no # è sbagliato #

WOZ: # è corretto? #

AN04: no è sbagliato

WOZ: qual è il numero della carta di credito?

AN04: quarantanove , settantasette cinquanta cinquantasei diciotto ventidue

6.4.7 Conclusioni

Abbiamo visto che la situazione di non comprensione da parte del sistema può attivare diverse strategie per la risoluzione dell'errore da parte dello stesso parlante. Seguendo l'approccio qualitativo di Fischer, si può vedere come i tentativi di risoluzione dell'errore da parte del parlante possano fornire informazioni su quello che il parlante pensa dell'enunciato appena formulato e delle capacità di comprensione della macchina. Oltre all'attenzione riservata alle tipologie di errore che occorrono generalmente con i sistemi di dialogo, nel capitolo ci siamo concentrati su come il parlante interpreta l'errore e su cosa ritiene più problematico nell'interazione con un sistema. La prima distinzione necessaria è quella tra ripetizioni e riformulazioni: in linea con i sistemi a iniziativa mista, la riformulazione è preferita alla ripetizione per quasi tutti i parlanti. L'ipotesi di errore che sta alla base

della ripetizione è un errore sul livello del canale o del riconoscimento; a cui vanno, rispettivamente, ascritte la ripetizione senza iperarticolazione e la ripetizione iperarticolata. Gli interventi lessicali, in particolar modo le sostituzioni, possono essere ricondotte a ipotesi sulle capacità del vocabolario del sistema. Gli interventi sintattici, invece, possono essere visti come una risposta ai difetti della grammatica di comprensione, a cui il parlante rimedia eliminando costruzioni sintattiche potenzialmente problematiche, e a un'incertezza su ciò che il sistema acquisito nei turni precedenti. Tale problematizzazione della coesione interna all'interazione, oltre a portare a frequenti espansioni sintattiche, è particolarmente evidente nel ripristino del referente di pronomi anaforici.

Al contrario, nei casi di errata acquisizione, in cui il sistema fornisce un feedback relativamente a ciò che ha compreso e di cui chiede conferma, la correzione è fortemente focalizzata su ciò che è errato della richiesta, a cui si associa nella maggior parte dei casi l'elisione di ciò che è stato compreso. Inoltre, il feedback porta a una maggiore presenza di ripetizioni; mentre le riformulazioni sono quasi sempre adattamenti alle strutture e al lessico usati dal sistema, che costituiscono un modello per il parlante. Nel *corpus* è ampiamente sfruttata l'iniziativa mista e la correzione diretta nel turno successivo alla richiesta di conferma: in questo è forte la distanza dal *corpus* C-ORAL-ROM, in cui l'iniziativa è spesso lasciata al sistema. Tuttavia, in un particolare caso di richiesta di conferma, relativo a una transazione, aumentano i casi in cui il parlante non prende l'iniziativa correggendo direttamente il sistema: dovrà essere indagata perciò, soprattutto su *corpora* di interazioni reali, la relazione tra gestione dell'iniziativa e importanza dell'attività o della sottoattività in svolgimento.

7. Conclusioni

La situazione comunicativa del dialogo uomo-macchina è circoscritta da parametri che ne riducono la variabilità e ne regolarizzano alcune caratteristiche: innanzitutto, le funzioni del sistema impongono alcune restrizioni, che possono rendere rilevanti o meno alcuni tratti dell'interazione (ad esempio, rendendo funzionali le sovrapposizioni o i gesti); poi, la stessa natura dell'interazione ("dialogo pratico", fortemente focalizzato sullo svolgimento di un'attività) lascia prevedere facilmente i principali contributi degli utenti, difficilmente estranei al dominio di applicazione, il quale, anzi, viene spesso esplicitato nel corso dell'interazione.

All'interno dei paletti posti dalla situazione comunicativa, lo svolgimento dell'interazione può essere parzialmente o totalmente condizionato dal tipo di iniziativa prevista dal sistema, che determina, da un punto di vista applicativo (sia per quel che riguarda la comprensione del sistema in un dato stato, sia per quel che riguarda la costruzione degli output), cosa un parlante può dire nel suo turno. Tuttavia, in linea con altri studi condotti sul sistema a iniziativa mista, nell'analisi del *corpus* C-ORAL-ROM abbiamo visto che la libertà concessa al parlante non si traduce automaticamente in minore controllo dell'attività da parte del sistema e che, anzi, lo spazio concesso non viene sempre sfruttato. Nel *corpus* C-ORAL-ROM, infatti, si possono distinguere due diversi poli di comportamento degli utenti, che vanno dal maggiore al minore controllo sullo svolgimento dell'interazione: suddividendo il *corpus* in due sotto-*corpus*, a seconda dell'iniziativa da parte degli utenti, si può notare una correlazione tra l'assenza di iniziativa e la scarsa varietà di elementi morfosintattici e costruzioni sintattiche, in favore di un "parlato telegrafico" (Danieli 2004) che presenta unicamente le variabili rilevanti per il sistema. Questo viene così usato in maniera affine a un'interfaccia grafica, senza che vi sia spazio per la negoziazione. Al contrario, i soggetti più attivi, che cercano, anche solo parzialmente, il controllo dell'interazione, sembrano essere maggiormente predisposti alla cooperazione, soprattutto nelle situazioni di errore.

I casi di errore sono quindi stati approfonditi nell'analisi del *corpus* Speaky, che presenta molte situazioni di errore appositamente simulate per studiare le strategie di risoluzione. Rispetto al *corpus* C-ORAL-ROM risalta l'ampia preferenza

per le riformulazioni rispetto alle ripetizioni: un sistema a iniziativa mista totale, stimolando un parlato più naturale meno incentrato sulle sole variabili sensibili per il sistema (i dati da fornire), porta il soggetto a variare la presentazione della richiesta o del comandi. L'analisi dei turni di riformulazione/ripetizione fornisce informazioni su quello che il parlante pensa del proprio input e delle capacità di comprensione della macchina. Nel corpus quello che sembra più preoccupare i parlanti sono errori di memoria del sistema o errori dovuti a un enunciato troppo pesante, mentre sembra meno problematico il livello lessicale (per quanto siano presenti sostituzioni) o il livello di riconoscimento vocale (l'iperarticolazione è solo marginale nel corpus). Nei casi di errata acquisizione il sistema fornisce invece un feedback relativamente a ciò che ha compreso e di cui chiede conferma: in questi casi, in cui il parlante ha informazioni su cosa è stato compreso dal sistema, la correzione è fortemente focalizzata sulla variabile errata presente nella richiesta di conferma e l'eliminazione del resto della richiesta. Nel *corpus* è inoltre ampiamente sfruttata la correzione diretta nel turno successivo alla richiesta di conferma, senza attendere la nuova richiesta del sistema: anche in questo caso il *corpus* Speaky si differenzia pesantemente dal *corpus* C-ORAL-ROM, caratterizzato spesso da risoluzioni guidate unicamente dal sistema. Tuttavia, si può notare un legame tra attesa di indicazioni del sistema e rilevanza dell'attività: pur essendo comunque preferita la correzione diretta, nel caso di conferma del numero di carta di credito aumentano i casi in cui il parlante non prende l'iniziativa correggendo direttamente il sistema e aspetta nuove istruzioni.

L'altro aspetto analizzato del *corpus* è stata la gestione del turno, a cui spesso corrisponde anche la gestione dell'iniziativa: l'analisi è stata condotta su alcuni stati del dialogo non regolati dalle domande del sistema. La prima indicazione rilevante è la non completa aderenza tra output del sistema e turno del sistema, dal punto di vista del parlante: nel caso dei saluti di apertura, ad esempio, alcuni soggetti salutano il sistema nella breve pausa che separa il saluto del sistema dalla richiesta di apertura. Nel caso dei saluti di chiusura, il turno successivo del parlante ai saluti del sistema è presente se nel saluto di chiusura viene vista una funzione rilevante (chiusura del canale comunicativo); nel caso sia giudicato irrilevante il saluto finale viene presto abbandonato dai soggetti che inizialmente lo usavano. La gestione dell'iniziativa può essere invece rilevata durante le transizioni delle attività, le quali possono avere configurazioni di turni diverse a seconda della collaborazione del parlante. Nella

maggior parte dei casi è il parlante a prendere il turno e aprire l'attività, velocizzando l'interazione; va considerato comunque un congruo numero di casi (per alcuni soggetti la principale strategia) in cui il parlante attende le indicazioni del sistema, o rimanendo in silenzio o fornendo un feedback di accettazione dell'azione che segnala inoltre il successo dell'interazione.

Anche se in alcuni casi non vi è completa aderenza tra output e turno del sistema, la maggior parte degli output del sistema sono seguiti da un breve silenzio, di media superiore a quello del dialogo tra persone, e caratterizzati da un'assenza di tentativi di sovrapposizione: il turno di parola concesso al sistema coincide con la lunghezza dell'output, senza la presenza di tentativi di presa del turno da parte del parlante. Tuttavia, durante alcuni output composti da più enunciati separati da una breve pausa (come ad esempio, il saluto iniziale o la richiesta di ripetizione) aumentano i tentativi di presa del turno, che possono comportare sovrapposizioni problematiche per il futuro sistema, nel caso verrà consentito il *barge-in*. Inoltre, il tempo di presa del turno può variare in maniera consistente sia tra parlante e parlante, sia a seconda dell'output del sistema.

Spostandoci su un piano applicativo, si possono identificare alcune situazioni potenzialmente problematiche, che andranno approfondite in futuro, e trarre alcune indicazioni generali. Innanzitutto gli sviluppatori devono confrontarsi con quella che ormai, in maniera riconosciuta anche in letteratura, viene vista come una polarizzazione di due tipi di comportamento. Per sistemi semplici, come il C-ORAL-ROM, la gestione dell'iniziativa da parte dell'utente può essere correlata alla varietà di forme e strategie usate dal parlante. L'identificazione di almeno due profili basati sull'iniziativa può consentire al sistema di variare le proprie caratteristiche a seconda del parlante: ad esempio, la possibilità di passare dall'iniziativa mista all'iniziativa singola (architettura generalmente più precisa e robusta) potrebbe migliorare la precisione del sistema con i parlanti che usano il sistema come una semplice interfaccia. Naturalmente, l'identificazione automatica del profilo dell'utente dovrà essere supportata da ulteriori studi, condotti su *corpora* più grandi, appositamente mirati sulla gestione dell'iniziativa nei dialoghi pratici e sulle eventuali correlazioni linguistiche e pragmatiche. Per quanto riguarda sistemi più aperti ed elastici, come la simulazione di Speaky, la polarizzazione si può invece notare nella gestione del turno, che differenzia gli utenti più attivi dagli utenti meno attivi, a seconda della percentuale di sequenze di turni iniziate dall'utente. Anche in questo caso, va

considerata la possibilità di dotare il sistema di due architetture diverse, che regolino il tempo di presa del turno del sistema a seconda del tipo di utente: per quelli meno attivi, ad esempio, sarebbe più pratico un sistema che non attenda un'eventuale intervento ma che guidi direttamente l'interazione senza attese che possono diventare frustranti. Per i tempi di presa del turno andrà in ogni caso approfondita la differenza che può intercorrere tra parlante e parlante, e tra tipo di output del sistema: abbiamo visto, infatti, che in alcune situazioni dialogiche il tempo di presa del turno si abbassa considerevolmente, “stimolando” addirittura sovrapposizioni (come nel caso di errori). Altre sovrapposizioni possono verificarsi quando il sistema fa attendere eccessivamente l'utente, il quale, invece di ripetere l'input, prova a completarlo: la distanza tra ciò che il parlante pensa di aver detto e ciò che effettivamente il sistema ha computato rischia di essere problematica per lo svolgimento dell'interazione.

In realtà, senza modificare le grammatiche, gran parte delle situazioni critiche, legate sia alla presa del turno sia alle richieste di ripetizione e conferma, possono essere risolte dal modulo di generazione del linguaggio, modulo che come abbiamo visto viene spesso considerato poco rilevante. Ad esempio, i possibili problemi che possono derivare nel caso di latenze del sistema possono essere risolti comunicando, con un feedback (anche ripetuto a distanze prefissate), che il sistema ha acquisito l'input e lo sta processando. Come abbiamo visto per le situazioni di incomprensione, per l'utente è rilevante sapere cosa il sistema sta processando e cosa aveva già precedentemente in memoria, per avere un appoggio su cui elaborare la propria riformulazione/ripetizione: anche quando il sistema non comprende un input, dovrebbe chiedere, insieme alla ripetizione, quali variabili ha già acquisito.

8. Bibliografia

ALEXANDERSSON ET AL. 1998 = J. Alexandersson, B. Buschbeck-Wolf, T. Fujinami, M. Kipp, S. Koch, E. Maier, N. Reithinger, M. Siegel, Dialogue Acts in Verbmobil 2, Verbmobil-Report Nr. 226, Saarbrücken: DFKI.

ALLEN ET AL. 2000 = J. F. Allen, D. Byron, M. Dzikovska, G. Ferguson, A. Galescu, A. Stent, An architecture for a generic dialogue shell, *Natural Language Engineering*, 6, pp. 1-16.

AMALBERTI ET AL. 1993 = R. Amalberti, N. Carbonell, P. Falzon, User representations of computer systems in human-computer speech interaction, *International Journal of Man-Machine Studies*, 38, pp. 547-566.

BALENTINE – MORGAN 1999 = B. Balentine, D.P. Morgan, *How to Build a Speech Recognition Application*, San Diego (California): Enterprise Integration Group, Inc.

BAMFORD 2000 = J. Bamford, *You can say that again. Repetition in discourse*, Bologna: Clueb.

BATLINER ET AL. 2006 = A. Batliner, C. Hacker, E. Noth, *Talk or not to Talk with a Computer: On-Talk vs. Off-Talk*, in K. Fischer (a cura di), *How People Talk to Computers, Robots, and Other Artificial Communication Partners*, SFB/TR 8 Report Spatial Cognition, Nr. 010-09/2006, University of Bremen, pp. 79–100.

BAZZANELLA 2002 = C. Bazzanella (a cura di), *Sul dialogo. Contesti e forme di interazione verbale*, Milano: Guerini.

BAZZANELLA 2005 = C. Bazzanella, Parlato dialogico e contesti di interazione, in K. Hölker, C. Maaß (a cura di), *Aspetti dell'italiano parlato*, Münster: LIT, pp. 1-22.

BAZZANELLA ET AL. 1999 = C. Bazzanella, G. Braiato, M. Danieli, The functions of repetition in human-machine dialogue, in M. Danieli, L. Lesmo (a cura di), *Atti del*

workshop AI*IA. Elaborazione del linguaggio e riconoscimento del parlato, IRST Trento, pp. 64-77.

BELL – GUSTAFSON 2000 = L. Bell, J. Gustafson, Positive and negative user feedback in a spoken dialogue corpus, in Proceedings of ICSLP '00 Beijing, pp. 589-592

BELL – GUSTAFSON 2003 = L. Bell, J. Gustafson, Child and adult speaker adaptation during error resolution in a publicly available spoken dialogue system, in Proceedings of Eurospeech 2003, Geneva, Switzerland, pp. 613-616.

BERINGER 2001 = N. Beringer, Evoking Gestures in SmartKom – Design of the Graphical User Interface, Report Nr. 3, Ludwig–Maximilians–Universität München, pp. 228-240.

BICKMORE – CASSELL 2005 = T. Bickmore, J. Cassell, Social Dialogue with Embodied Conversational Agents, in J. van Kuppevelt, L. Dybkjaer, N. Bernsen (a cura di), *Advances in Natural, Multimodal Dialogue Systems*, 30, New York: Kluwer Academic, pp. 23-54.

BOBROW ET AL. 1977 = D. G. Bobrow, R. M. Kaplan, M. Kay, D. A. Norman, H. Thompson, T. Winograd, GUS, A frame driven dialog system, *Artificial Intelligence*, 8, pp. 155–173.

BOHUS – RUDNICKY 2005 = D. Bohus, A. Rudnicky, Sorry, I didn't catch that! An investigation of non-understanding errors and recovery strategies, in Proceedings of SIGDial 2005, pp. 2781-2784.

BOYCE 2008 = S. J. Boyce, Spoken natural dialog systems: User interface issues for the future, in D. Gardner-Bonneau, H. E. Blanchard (a cura di), *Human factors and voice interactive systems*, Norwell: Kluwer Academic, pp. 43-80

BRANIGAN ET AL. 2003 = H. Branigan, M. Pickering, J. Pearson, J. McLean, C. Nass, Syntactic alignment between computers and people: the role of belief about mental states, in Proceedings of CogSc, pp. 186-191.

BRANIGAN ET AL. 2006 = H. Branigan, J. Pearson, H. P. Hu, M. J. Pickering, C. Nass, Adaptive language behavior in HCI: how expectations and beliefs about a system affect users' word choice, in Proceedings of the SIGCHI conference on Human Factors in computing systems, pp.1177-1180.

BRENNAN 1998 = S. Brennan, The grounding problem in conversations with and through computers, in S. R. Fussell, R. Kreuz (a cura di), *Social and Cognitive Psychological Approaches to Interpersonal Communication*, Hillsdale: Lawrence Erlbaum Associates, pp. 201–225.

BUNT 1994 = H. Bunt, *Context and dialogue control*, Think, 3, pp. 19–31.

BUNT 2000 = H. Bunt, Dynamic interpretation and dialogue theory, in M. M. Taylor, F. Neel, D. G. Bouwhuis (a cura di), *The structure of multimodal dialogue*, Amsterdam: John Benjamins, pp. 139–166.

BURKE 2004 = P. Burke, Segmentation and control of a dissertation defense, in A. Grimshaw (a cura di), *What's going on here? Complementary studies of talk*, Norwood: Ablex, pp. 95-124.

CARLETTA ET AL. 1997 = J. Carletta, A. Isard, S. Isard, J. C. Kowtko, G Doherty-Sneddon, A. H. Anderson, The reliability of a dialogue structure coding scheme, *Computational Linguistics*, 23 (1), pp. 13–32

CASELL 2009 = J. Cassell, Dialogue and conversational agents, in JURAFSKY – MARTIN 2009, pp. 847-894.

CERRATO 2006 = L. Cerrato, *Sulle tecniche di elicitazione di parlato semi-spontaneo*, Progetto Clips, disponibile su http://www.clips.unina.it/it/documenti/2_tecniche_di_elicitazione_dialogica.pdf

CHU-CARROLL – BROWN 1997 = J. Chu-Carroll, M. K. Brown, Tracking initiative in collaborative dialogue interactions, in Proceeding ACL/EACL-97, Madrid, pp. 262–

CLARK 1996 = H. H. Clark, , *Using Language*. Cambridge: Cambridge University Press.

CLARK – SCHAEFER 1989 = H. H. Clark, E. F. Schaefer, Contributing to discourse, *Cognitive Science*, 13, pp. 259-294.

COHEN ET AL. 1999 = Cohen, R., Allaby, C., Cumbaa, C., Fitzgerald, M., Ho, K., Hui, B., Latulipe, C., Lu, F., Moussa, N., Pooley, D., Qian, A. and S. Siddiqi; What is Initiative?; in S. Haller, S. McRoy, A. Kobsa (a cura di), *Computational Models of Mixed-Initiative Interaction*, Norwell: Kluwer Academic Publishers.

COHEN ET AL. 2004 = M. H. Cohen, J. P. Giangola, J. Balogh, *Voice User Interface Design*, Boston: Addison-Wesley

CRESTI ET AL. 2004: E. Cresti, M. Moneglia, A. Moreno Sandoval, J. Veronis, M. F. Bacelar, *C-ORAL-ROM. Integrated Reference Corpora for Spoken Romance Languages*, Amsterdam / Philadelphia: John Benjamins.

DAHLBÄCK 1991 = N. Dahlback, *Representations of Discourse: Cognitive and Computational Aspects*, Ph.D. thesis, Linköping University.

DAHLBÄCK 1992 = N. Dahlbäck, Pronoun usage in NLI-dialogues: A Wizard of Oz study, in Papers from the Third Nordic Conference on Text Comprehension in Man and Machine, Linköping, pp. 27-42.

DAHLBÄCK – JÖNSSON 1992 = N. Dahlbäck, A. Jönsson, An empirically based computationally tractable dialogue model, in Proceedings of the Fourteenth Annual Meeting of The CognitiveScience Society, Bloomington, pp. 785-790.

DAHLBÄCK ET AL. 1993 = N. Dahlbäck , A. Jönsson , L. Ahrenberg, Wizard Of Oz Studies - Why And How, in Proceedings of the 1st international conference on

Intelligent user interfaces, pp. 193-200.

DAMIANO 2002 = R. Damiano, Modelli teorici del dialogo persona-macchina, in BAZZANELLA 2002, pp. 223-235.

DANIELI 2002 = M. Danieli, Il dialogo persona-macchina: applicazioni, in BAZZANELLA 2002, pp. 237-251.

DANIELI 2004 = M. Danieli, Il parlato telegrafico tra persone e sistemi artificiali, in F. Albano Leoni, F. Cutugno, M. Pettorino, R. Savy (a cura di), *Il parlato italiano*, Atti del Convegno nazionale di Napoli, Napoli: M. D'Auria Editore

DANIELI – GERBINO 1995 = M. Danieli, E. Gerbino, Metrics for evaluating dialogue strategies in a spoken language system, in Proceedings of the 1995 AAAI Spring Symposium on Empirical Methods in Discourse Interpretation and Generation, Cambridge, pp. 34–39

DE MORI 1998 = R. De Mori (a cura di), *Spoken dialogues with computers*, London: Elsevier.

DE MORI ET AL. 2008 = R. De Mori, F. Bechet, D. Hakkani-Tur, M. McTear, G. Riccardi, G. Tur, Spoken Language Understanding: A Survey, *IEEE Signal Processing Magazine*, 25, pp. 50-58.

DUNCAN 1972 = S. Duncan, Some signals and rules for taking speaking turns in conversations, *Journal of Personality and Social Psychology*, 23, pp. 283–292.

DURANTI 1986 = A. Duranti, The Audience as Co-Author: An Introduction, *Text*, 6, pp. 239-47.

EAGLES 1998 = G. Leech, M. Weisser, A. Wilson, M. Grice, *LE-EAGLES-WP4-4 Integrated Resources Working Group Survey and guidelines for the representation and annotation of dialogue*, disponibile su <http://coral.lili.uni-bielefeld.de/EAGLES/WP4/wp4apr5/>

EDLUND ET AL. 2006 = J. Edlund, J. Gustafson, M. Heldner, Two faces of spoken dialogue systems, in Interspeech 2006 - ICSLP Satellite Workshop Dialogue on Dialogues: Multidisciplinary Evaluation of Advanced Speech-based Interactive Systems, Pittsburgh, pp. 51-54.

EDLUND ET AL. 2008 = J. Edlund, J. Gustafson, M. Heldner A. Hjalmarsson, Towards human-like spoken dialogue systems, *Speech communication*, 50, pp. 630-645.

FISCHER – WREDE 1997 = K. Fischer, B. Wrede, Discourse particles in female and male human-computer-interaction, in R. Lander, A. Adam (a cura di), *Women into Computing*, Bristol: Intellect, pp. 36-49.

FISCHER 2003 = K. Fischer, Linguistic methods for investigating concepts in use, in T. Stolzi, K. Kolbe (a cura di), in *Methodologie in der Linguistik*, Frankfurt: Peter Lang, pp. 39-62.

FISCHER 2006 = K. Fischer, The Role of Users' Preconceptions in Talking to Computers and Robots, in K. Fischer (a cura di), Proceedings of the Workshop on How People Talk to Computers, Robots, and other Artificial Communication, Delmenhorst, Germany, pp. 112-130.

FISCHER 2006B = K. Fischer, *What Computer Talk Is and Isn't*, Saarbrücken: AQ-Verlag.

FISCHER – BATEMAN 2006 = K. Fischer, J. A. Bateman, Keeping the initiative: an empirically motivated approach to predicting user-initiated dialogue contributions in HCI, in Proceedings of EACL '06, pp. 185–192.

FISCHER – LOHSE 2007 = K. Fischer, M. Lohse, Shaping naive users' models of robots' situation awareness, in Proceedings of the 16th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN 2007), Jeju, Korea, pp. 534 – 539.

FORD – THOMPSON 1996 = C. Ford, S. Thompson, Interactional units in conversation: syntactic intonational and pragmatic resources for the management of turns, in E. Ochs, E. Schegloff, S. Thompson (a cura di), *Interaction and Grammar*, Cambridge University Press, pp. 134–184.

FRASER – GILBERT 1991 = N. Fraser, N. Gilbert, Simulating speech systems, *Computer Speech and Language*, 5, pp. 81-99.

GAL 1988 = A. Gal, *Cooperative Responses in Deductive Databases*, PhD thesis, Department of Computer Science, University of Maryland, College Park.

GIESELMANN – WAIBEL 2005 = P. Gieselmann, A. Waibel, What makes human-robot dialogues struggle?, in Proceedings of the Ninth Workshop on the Semantics and Pragmatics of Dialogue 2005, Nancy, pp. 100-108.

GONZALEZ-LEDESMA – MORENO-SANDOVAL 2005 = A. González-Ledesma, A. Moreno-Sandoval, in Pragmatic Analysis of Man-Machine Interactions in a Spontaneous Speech Corpus, in P. Danielsson, M. Wagenmakers (a cura di), Proceedings Corpus Linguistic Conference, vol. 1, Birmingham, 2005, pp. 220-232.

GORIN ET AL. 1997 = A.L. Gorin, G. Riccardi, J.H. Wright, How may I Help You?, *Speech Communication*, 23, pp. 113-127.

GRAVANO – HIRSCHBERG 2011 = A. Gravano, J. Hirschberg, Turn-taking cues in task-oriented dialogue, *Computer Speech and Language*, 25 (3), pp. 601-634.

GREEN ET AL. 2004 = A. Green, Helge Hottenrauche, K. Eklund, K. Severinson: Applying the Wizard of Oz Framework to Cooperative Service discovery and Configuration, in ROMAN 2004 13th IEEE International Workshop on Robot and Human Interactive Communication, Okayama, Japan, pp. 575-580.

GROSZ 1977 = B. J. Grosz, *The Representation and Use of Focus in Dialogue Understanding*, Unpublished Ph.D. Thesis, University of California, Berkeley.

GUINDON 1988 = R. Guindon, A multidisciplinary perspective on dialogue structure in user-adviser dialogues, in R. Guindon (a cura di), *Cognitive science and its applications for human-computer interaction*, Hillsdale: Lawrence Erlbaum Associates, pp. 163-200.

GUINDON ET AL. 1986 = R. Guindon, P. Sladky, H. Brunner, J. Conner, The Structure of User-Adviser Dialogues: Is There Method in Their Madness?, in Proceedings of 24th Meeting of the Association for Computational Linguistics ACL86, New York, pp. 224-230.

GUINDON ET AL. 1987 = R. Guindon, K. Shuldborg, J. Conner, Grammatical and ungrammatical structures in user-adviser dialogues: evidence for sufficiency of restricted languages in natural language interfaces to advisory systems, in Proceedings of 25th Meeting of the Association for Computational Linguistics ACL87, Stanford, pp. 41-44.

GUSTAFSON ET AL. 1997 = J. Gustafson, A. Larsson, R. Carlson, K. Hellman, How do system questions influence lexical choices in user answers?, in Proceedings of Eurospeech '97, pp. 2275-2278.

GUSTAFSON – BELL 2000 = J. Gustafson, L. Bell, Speech technology on trial - Experiences from the August system, *Natural Language Engineering*, 6, pp. 273-286.

GUSTAFSON 2002 = J. Gustafson, *Developing Multimodal Spoken Dialogue Systems. Empirical Studies of Spoken Human-Computer Interaction*, Ph. D. thesis, KTH, Stockholm.

GUSTAFSON ET AL. 2008 = J. Gustafson, M. Heldner, J. Edlund, *Potential benefits of human-like dialogue behaviour in the call routing domain*, in Perception in multimodal dialogue systems, Berlin, Heidelberg: Springer, pp. 240–251

GUYOMARD – SIROUX 1987 = M. Guyomard, J. Siroux, Experimentation in the specification of an oral dialogue, in Proceedings of the NATO Advanced Study

Institute on Recent advances in speech understanding and dialog systems, pp. 497-501.

HELDNER – EDLUND 2010 = M. Heldner, J. Edlund, Pauses, gaps and overlaps in conversations, *Journal of Phonetics*, 38, 555-568.

HITZENBERGER – WHOMSER-HACKER 1995 = L. Hitzenberger, C. Womser-Hacker, Experimentelle Untersuchungen zu multimodalen natürlichsprachigen Dialogen in der Mensch-Computer-Interaktion, *Sprache und Datenverarbeitung*, 19(1), 51–61.

KELLEY 1984 = J. F. Kelley, An iterative design methodology for user-friendly natural language office information applications, *ACM Transactions on Information Systems*, 2 (1), pp. 26-41.

KRAUSE 1992 = J. Krause, Fazit und Ausblick: Registermodell versus metaphorischer Gebrauch von Sprache in der Mensch-Computer- Interaktion, in KRAUSE – HITZENBERGER 1992, pp. 157-170.

KRAUSE – HITZENBERGER 1992 = J. Krause, L. Hitzenberger (a cura di), *Computer Talk*, Hildesheim, Olms Verlag.

JEKAT ET AL. 1995 = S. Jekat, A. Klein, E. Maier, I. Maleck, M. Mast, J. Quantz, Dialogue Acts in VERBMOBIL, Verbmobil Report Nr. 65, Saarbrücken: DFKI.

JOHNSTON ET AL. 1994 = A. Johnstone, U. Berry, T. Ngyuen, A. Asper, There was a long pause: Influencing turn-taking behaviour in human-human and human-computer spoken dialogues, *International Journal of Human-Computer Studies*, 41, pp. 383-411.

JÖNSSON – DAHLBÄCK 1988 = A. Jönsson, N. Dahlbäck, Talking to a Computer Is Not like Talking to Your Best Friend, in Proceedings of SCAI, Tromsø, pp. 53-68.

JURAFSKY – MARTIN 2009 = D. Jurafsky, J. H. Martin, *Speech and language processing*, Upper Saddle River, New Jersey: Pearson.

KELLEY 1984 = J. F. Kelley, An iterative design methodology for user-friendly natural language office information applications, *ACM Transactions on Information Systems*, 2 (1), pp. 26-41.

KRAHMER ET AL. 1999 = E. Krahmer, M. Swerts, M. Theune, M Weegels, Problem spotting in human-machine interaction, in Proceedings of Eurospeech 1999, Budapest, Hungary, pp. 1423-1426.

KRAHMER ET AL. 2001 = E. Krahmer, M. Swerts, M. Theune, M Weegels, Error detection in spoken human-machine interaction, *International Journal of Speech Technology*, 4, pp.19-30.

LARSSON 2003 = S. Larsson, Interactive communication mangement in an issue-based dialogue system, in Proceedings of the 7th Workshop on the semantics and pragmatics of dialogue, Saarbrücken, Germany, pp. 75-82.

LEECH – WEISSER 2005 = G. Leech, M. Weisser, *Pragmatics and dialogue*, in R. Mitkov (a cura di), *The Oxford Handbook of Computational Linguistics*, Oxford University Press, pp. 136-156

LERNER 2003 = G. H. Lerner, Selecting next speaker: The context-sensitive operation of a context-free organization, *Language in Society*, 32(2), pp. 177-201

LEVOW 1998 = G. A. Levow, Characterizing and recognizing spoken corrections in human-computer dialogue, in Proceedings of Coling/ACL '98 , pp. 736-742.

McPHEAT 2013 = N. McPheat, Apple has “Oversold Siri” says 46% of Americans, *Intelligent Voice*, disponibile su <http://www.intelligentvoice.com/blog/new-poll-apple-oversold-siri-says-46-americans/>

MOREL 1986 = M. A. Morel, *Computer-human interaction*, in Nato research study group on ASP and CHI in command and control: structures of multimodal dialogue including voice, Venaco, France.

MOREL 1989 = M.-A. Morel, Computer-human communication, in M. Taylor, F. Neel, D. Bouhuis (a cura di), *The Structure of Multimodal Communication*, Amsterdam: North-Holland Elsevier, pp. 323–330.

MORTARA GARAVELLI 1979 = B. Mortara Garavelli, *Il filo del discorso*, Torino: Giappichelli.

NASS – MOON 2000 = C. Nass, Y. Moon, Machines and mindlessness: Social responses to computers, *Journal of Social Issues*, 56(1), pp. 81–103.

NASS – BRAVE 2005 = C. Nass, S. Brave, *Wired for Speech. How Voice Activates and Advances the Human-Computer Relationship*, Cambridge: MIT Press.

NORMAN 1988 = D.A. Norman, *The Design of Everyday Things*, New York: Basic Books.

NOVICK – SUTTON 1997 = D. Novick, S. Sutton, What is mixed-initiative interaction?, in *Papers from the 1997 AAAI Spring Symposium on Computational Models for Mixed Initiative Interaction*, Stanford University, pp. 93-102.

OGDEN 1988 = W. C. Ogden, Using Natural Language Interfaces, in M. Helander, *Handbook of Human-Computer Interaction*, New York: North Holland, pp. 205-235.

OVIATT 1995 = S. Oviatt, Predicting spoken disfluencies during human-computer interaction, *Computer Speech and Language*, 9, pp. 19–35

OVIATT ET AL. 1998 = S. Oviatt, J. Bernard, G. A. Levow, Linguistic adaptations during spoken and multimodal error resolution, *Language and Speech*, 41, pp. 419-442.

PAEK 2003 = T. Paek, Toward a taxonomy of communication errors, in *Proceedings of ISCA Tutorial and Research Workshop on Error Handling in Spoken Dialogue Systems*, Château d'Oex, Switzerland, pp. 53-58.

PARENT – ESKENAZI 2010 = G. Parent, M. Eskenazi, Lexical entrainment of real users in the let's go spoken dialog system, in Proceedings of ISCA Interspeech 2010, pp. 3018-3021

PÄTZOLD ET AL. 1995 = M. Pätzold, M. Scheffers, A.P. Simpson, W. Thon, Controlled elicitation and processing of spontaneous speech in Verbmobil, in Proceedings of the International Congress of Phonetic Sciences, Stockholm, pp. 314-317.

PIERACCINI 2012 = R. Pieraccini, *The Voice in the Machine. Building Computers That Understand Speech*, Cambridge: MIT Press

PIERACCINI ET AL. 1991 = R. Pieraccini, E. Levin, C. H. Lee, Stochastic representation of conceptual structure in the ATIS task, in Proceedings DARPA Speech and Natural Language Workshop, Pacific Grove, CA, pp. 121–124.

PORZEL – BAUDIS 2004 = R. Porzel, M. Baudis, The tao of chi: Towards effective human-computer interaction, in Proceedings of HLT-NAACL, pp. 209–216.

QVARFORDT 2004 = P. Qvarfordt, *Eyes on multimodal interaction*, Doctoral dissertation, Linköping University.

RAUX – ESKENAZI 2008 = A. Raux, M. Eskenazi, *Optimizing Endpointing Thresholds using Dialogue Features in a Spoken Dialogue System*, in Proceedings SIGdial 2008, Columbus, OH

REEVES – NASS 1996 = B. Reeves, C. Nass, *The Media Equation*, Stanford: CSLI and Cambridge: Cambridge University Press.

REILLY 1987 = R. Reilly, Ill-formedness and miscommunication in person-machine dialogue, *Information and Software Technology*, 29(2), pp. 69-74.

REITER 2010 = E. Reiter, Natural Language Generation, in A. Clark, C. Fox, S. Lappin (a cura di), *Handbook of Computational Linguistics and Natural Language Processing*, pp. 574-598.

REITER – DALE 2000 = E. Reiter, R. Dale, *Building Natural-Language Generation Systems*, Cambridge: Cambridge University Press.

RICCARDI – GORIN 2000 = G. Riccardi, A. L. Gorin, Stochastic Language Adaptation over Time and State in Natural Spoken Dialogue Systems, *IEEE Transactions on Speech and Audio Processing*, 8, pp. 3-10.

RICHARDS – UNDERWOOD 1984a = M. Richards, K. Underwood, Talking to machines: How are people naturally inclined to speak?, in Proceedings of the Ergonomics Society Annual Conference, pp. 67-76.

RICHARDS – UNDERWOOD 1984b = M. Richards, K. Underwood, How should people and computers speak to each other?, in Proceedings of Interact '84, pp. 33-36.

RIEK 2012 = L. D. Riek, Wizard of Oz Studies in HRI: A Systematic Review and New Reporting Guidelines, *Journal of Human Robot Interaction*, 1, pp. 119-136

SAVY 2010 = R. Savy, Pr.A.T.I.D: a coding scheme for pragmatic annotation of dialogues, in Proceedings of LREC 2010, Malta, pp. 2141-2148.

SACKS ET AL. 1974 = H. Sacks, E. Schegloff, G. Jefferson, A simplest systematics for the organization of turn-taking for conversation, *Language*, 50, pp. 696-735.

SAYGIN – CICEKLI 2002 = A.P. Saygin, I. Cicekli, Pragmatics in human-computer conversation, *Journal of Pragmatics*, 34(3), pp. 227-258.

SCHEGLOFF – SACKS 1973 = E. A. Schegloff, H. Sacks, Opening up closings, *Semiotica*, 8, pp. 289-327.

SCHLANGEN 2004 = D. Schlangen, *Causes and strategies for requesting clarification in dialogue*, in Proceedings of SIGdial 2004, Boston, USA, pp. 136-143.

SENEFF – POLIFRONI 2000 = S. Seneff, J. Polifroni, *Dialogue management in the*

mercury flight reservation system, in ANLP/NAACL Workshop on Conversational Systems, Seattle, pp. 11-16.

SELTING 2000 = M. Selting, The construction of units in conversational talk, *Language in Society*, 29, pp. 477-517.

SINGH ET AL. 2002 = S. Singh, D. J. Litman, M. J. Kearns, M. A. Walker, Optimizing dialogue management with reinforcement learning: Experiments with the njfun system, *Journal of Artificial Intelligence*, 16, pp. 105–133.

SI-TAL = Dialoghi Annotati per Interfacce vocali avanzate – Specifiche Tecniche, disponibile su http://www.ilc.cnr.it/dialoghiannotati_prg/papers/DialoghiAnnotati-Linea1.1-SpecificheTecniche.pdf

SORNICOLA 1981 = R. Sornicola, *Sul parlato*, Bologna: Il Mulino.

STALNAKER 1978 = R. C. Stalnaker, Assertion, in P. Cole (a cura di), *Pragmatics: Syntax and Semantics*, 9, pp. 315–332.

STOYANCHEV – STENT 2009 = S. Stoyanchev, A. Stent, Lexical and Syntactic Adaptation and Their Impact in Deployed Spoken Dialog Systems, in *Proceedings of HLT-NAACL 2009*, pp.189-192

SWERTS ET AL. 2000 = M. Swerts, D. J. Litman, J. Hirschberg, Corrections in spoken dialogue systems, in *ICSLP-00*, Beijing, China, pp. 615-618

TRAUM 2000 = D. Traum, 20 questions for dialogue act taxonomies, *Journal of Semantics*, 17(1), pp. 5-30

TRAUM – HINKELMAN 1992 = D. Traum, E. A. Hinkelman, Conversation acts in task-oriented spoken dialogue, *Computational Intelligence: Special Issue on Computational Approaches to Non-Literal Language*, 8(3), pp. 575–599

TRAUM – LARSSON 2003 = D. Traum, S. Larsson, The information state approach to

dialogue management, in J. van Kuppevelt, R. Smith (a cura di), *Current and New Directions in Discourse and Dialogue*, Norwell: Kluwer Academic, pp. 325-353

VARILE – ZAMPOLLI 1997 = G. B. Varile, A. Zampolli (a cura di), *Survey of the state of the art in human language technology*, Pisa

WALKER – WHITTAKER 1990 = M. A. Walker, S. Whittaker, Mixed initiative in dialogue: An investigation into discourse segmentation, in Proceedings of the 28th Meeting of the ACL, pp. 70-78.

WALKER – RAMBOW 2002 = M. A. Walker, O. C. Rambow, Spoken language generation, *Computer Speech and Language. Special Issue on Spoken Language Generation*, 16 (3-4), pp. 273–281

WARD – HEEMAN 2000 = K. Ward, P. A. Heeman, Acknowledgements in human-computer interaction, in Proceedings of the ANLP-NAACL, pp. 281–287

WARD ET AL. 2005 = N. Ward, A. Rivera, K. Ward, D. Novick, Root causes of lost time and user stress in a simple dialog system, in Proceedings of the 9th European Conference on Speech Communication and Technology (Interspeech 2005-Eurospeech), Lisbon, pp. 1565-1568.

WEI-HAAS ET AL. 1985 = L. Wei-Hass, P. Green, The Rapid Development of User Interfaces: Experience with the Wizard of OZ Method, in Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 29 (5), pp. 470–474

WEINSCHENK – BARKER 2000 = S. Weinschenk, D. T. Barker, *Designing effective speech interfaces*, Hoboken: Wiley.

WEIZENBAUM 1966 = J. Weizenbaum, ELIZA – A Computer Program For the Study of Natural Language Communication Between Man and Machine, *Communications of the ACM*, 9 (1), pp. 36-45

WIREN ET AL. 2007 = M. Wiren, R. Eklund, F. Engberg, J. Westermarck, Experiences

of an in-service Wizard-of-Oz data collection for the deployment of a call-routing application, in Proceedings of the Workshop on Bridging the Gap: Academic and Industrial Research in Dialog Technologies, pp. 56-63

YOUNG 2010 = S. Young, Frederick Jelinek 1932 – 2010 : The Pioneer of Speech Recognition Technology, Speech and Language Processing Technical Committee Newsletter (IEEE Signal Processing Society), disponibile su <http://www.signalprocessingsociety.org/technical-committees/list/sl-tc/spl-nl/2010-11/jelinek/>